ICSIIT

2nd International Conference on Soft Computing, Intelligent System and Information Technology 2010

Proceedings of 2nd ICSIIT 2010

International Conference on Soft Computing, Intelligent System and Information Technology

1-2 July 2010, Bali, Indonesia

ICSIIT

Supported by

MERATUS

APTIKOM

ASOSIASI PERGURUAN TINGGI INFORMATIKA DAN KOMPUTER

IBM

Proceedings

ICSIIT 2010

International Conference on
Soft Computing, Intelligent System and Information Technology

1-2 July 2010

Bali, Indonesia

Editors:
Leo Willyanto Santoso
Andreas Handojo

Informatics Engineering Department
Petra Christian University

Center of Soft Computing and
Intelligent System Studies

Proceedings

International Conference on Soft Computing, Intelligent System and Information Technology 2010

Publisher:
Informatics Engineering Department
Petra Christian University

ISBN: 978-602-97124-0-7

Additional copies may be ordered from:
Informatics Engineering Department
Petra Christian University, Siwalankerto 121-131, Surabaya 60236, Indonesia

Cover Art production by Adi Wibowo/Informatics Engineering Department

ICSIIT 2010
Table of Contents

Artificial Intelligence & Applications

Knowledge & Data Engineering

Imaging Technology

Internet, Web Services & Mobile Applications

Communication Systems & Networks

Preface

First of all, I would like to give thank to God the Creator, God the Redeemer and God who leads us to the truth for all His blessings to us. As we all know, this 2nd International Conference on Soft Computing, Intelligent Systems and Information Technology 2010 (ICSIIT 2010) is held from 1-2 July 2010 in the Hard Rock Hotel located at this paradise island, Bali, Indonesia. I thank Him for His presence and guidance in letting this conference happen. Only by God's grace, we hope we could give our best for 2nd ICSIIT 2010 despite of all of our limitation.

We have received more than 130 papers from 15 countries. Only 96 papers from 13 countries have been accepted based on reviewers' ratings and comments. The paper selection process was based on full paper submissions. We thank all authors who have contributed and participated in presenting their works at this conference. We also gratefully acknowledge the important review supports provided by the 19 members of the program committee from 8 different countries. Their efforts were crucial to the success of the conference.

We are also so blessed by the presence of two invited speakers who will address the important trends relating to natural languages processing and soft computing. The first issue on natural language will be addressed by a lovely professor, Prof. Rachel Edita O. Roxas, Phd. who will present "Human Language Technology: the Philippine Context". We are aware that the main problem in language processing is ambiguity from syntax level to semantic level. In my personal opinion, we are also living in between inherently ambiguous and completely reasonable world. Einstein once said that "As far as the laws of mathematics refer to reality, they are not certain, as far as they are certain, they do not refer to reality." Prof. Rolly Intan, Dr.Eng will address this issue on soft computing with his presentation entitled "Mining Multidimensional Fuzzy Association Rules from a Normalized Relational Database".

I hope during your stay in this beautiful island you will enjoy and benefit both, the fresh sea breeze and harmonious sound from sea waves, as well as the intellectual and scientific discussions. I hope your contributions and participation of the discussion will lead to the benefit of the advancements on Soft Computing, Intelligent Systems and Information Technology.

Soli Deo Gloria,
Iwan Njoto Sandjaja
Conference Chair
ICSIIT 2010 Bali Indonesia

Organizing Committee

The first ICSIIT 2010 is organized by Informatics Engineering Department, in cooperation with the Center of Soft Computing and Intelligent System Studies, Petra Christian University, Indonesia.

Conference Chair:

Iwan Njoto Sandjaja	Petra Christian University, Indonesia
Adi Wibowo	Petra Christian University, Indonesia

Organizing Committee:

Agustinus Noertjahyana	Petra Christian University, Indonesia
Alexander Setiawan	Petra Christian University, Indonesia
Andreas Handojo	Petra Christian University, Indonesia
Djoni Haryadi Setiabudi	Petra Christian University, Indonesia
Gregorius Satia Budhi	Petra Christian University, Indonesia
Ibnu Gunawan	Petra Christian University, Indonesia
Justinus Andjarwirawan	Petra Christian University, Indonesia
Kartika Gunadi	Petra Christian University, Indonesia
Leo Willyanto Santoso	Petra Christian University, Indonesia
Liliana	Petra Christian University, Indonesia
Lily Puspa Dewi	Petra Christian University, Indonesia
Rudy Adipranata	Petra Christian University, Indonesia
Silvia Rostianingsih	Petra Christian University, Indonesia
Yulia	Petra Christian University, Indonesia

Program Committee

Grant Pogosyan (Japan)

Jan Chan (Australia)

Kevin Wong (Australia)

Kwan Pyo, Ko (Korea)

Masao Mukaidono (Japan)

Moeljono Widjaja (Indonesia)

M. Rahmat Widyanto (Indonesia)

Nelson Marcos (Philippines)

Masashi Emoto (Japan)

Noboru Takagi (Japan)

Rolly Intan (Indonesia)

Budi Bambang (Indonesia)

Rudy Setiono (Singapore)

Shan Ling, Pan (Singapore)

Son Kuswadi (Indonesia)

Tae Soo, Yun (Korea)

Xiuying Wang (Australia)

Yung Chen, Hung (Taiwan)

Zuwairie Ibrahim (Malaysia)

Human Language Technology: The Philippine Context

Rachel Edita Roxas
Center for Language Technologies
College of Computer Studies
De La Salle University
rachel.roxas@delasalle.ph

Allan Borra
Center for Language Technologies
College of Computer Studies
De La Salle University
borgz.borra@delasalle.ph

ABSTRACT

We present the diverse research activities on human language technology considering the Philippine context: its geography, history and people. These projects include the formal representation of human languages and the processes involving these languages cutting across various forms such as text, speech and video files. Both rule-based and example-based approaches have been used in various experiments and have shown to be complementary in computation scenarios. Applications on languages that we have worked on include Machine Translation, Natural Language Generation, Information Extraction, and Audio and Video Processing. These applications provide the current human language interface for communication, searching, and learning, to name a few.

Keywords

Human Language Technology, Natural Language Processing

1. INTRODUCTION

Human language or natural language is a means of communication between and among human beings. Human languages are non-static and evolving from various times and places. The use of human language is pervasive; it can be used in discourses when communication is personal or over distances through various media such as telephone or the internet. Human language can take on various forms from textual mode, to audio and even video to capture non-verbal cues such as those used for sign languages. As the saying goes, "Listen to what I am not saying" pertains to non-verbal communication through gestures and cues that allow humans to interpret what is being said appropriately through what is not being said.

Human language is culture; and understanding human language allows us to understand our culture, who we are as communities and nations, and as human beings. It embodies an intertwine of social issues coupled with a sense of identity. In the 1800s, Ornolfor Thorsson, an adviser of the President of Iceland, at the time when Icelandic language was in danger of disappearing after years of Norwegian colonialism, said, "Without our language, we have no culture, we have no identity, we are nothing." National change can also be effected through language. In the previous century, our national hero Jose Rizal said "the pen is mightier than the sword."

There are more than 6,000 living human languages in the world. In the Philippines alone, there are 168 natively spoken languages [32] spread across the 7,100 islands of the archipelago. Aside from these Philippine languages, there are dialects or variants from locality to locality. Also, because of past colonial influences, the use of non-indigenous languages (such as English, Spanish and Chinese) can also be found in Philippine society now, and linguistic studies are also focused on understanding the linguistic phenomena in the use of these foreign languages in the Philippine context. For instance, English is still being used as the medium of instruction in our schools, although there are new advocacies to using the mother tongue in our schools in the early years of schooling.

Unfortunately, today, the Philippines has one of the highest rates of dying languages in the world (Solfed Foundation Inc). Thus, there is an advocacy to preserve our languages especially those that are already near extinction and almost dying. This can be done through documentation of our languages through print, audio and video forms. Unfortunately at this time, most of the available documentation have been focused on the major languages of the country.

Aside from documentation of these languages, the study of languages is a step further. Linguistics studies on Philippine languages have also been more focused on major languages, specifically, Tagalog and Ilocano (Liao, 2006).

Although many universities in the Philippines are known for research on applied linguistics, relatively few works have been done on the acquisition and/or processing of Philippine languages. And these will be the focus of this paper, specifically on language documentation and processing using technology, or called human language technologies.

Language tools are applications that support linguistic research on language resources and processing for various language computational layers. These include lexical units, to syntax and semantics. These language resources and processes usually employ either a rule-based approach or an example-based approach. In general, rule-based approaches formally capture language resources and processes which would require consultations and inputs from linguists. On the other hand, example-based approaches employ methodologies where automatic learning of rules is performed based on examples that are fed into the system, some of which are non-linguistic in processing behavior.

These two general approaches are not mutually-exclusive but are complementary in nature. Rules result to robustness, while

examples allow for variances and deviations from the linguistically-accepted rules.

2. LANGUAGE RESOURCES

We discuss our experiments on various technologies for automatic (or semi-automatic) extraction of language resources such as the lexicon, morphological information, grammar, and the corpora In the process, work has been focused on manual construction of these language resources to be used as seed data for our automatic and semi-automatic approaches. This is to formally capture the intricacies of Philippine languages, designed with the intention of using them for various language technology applications. These materials were painstakingly worked on since we had to literally build them from almost non-existent digital forms.

One of the main resources of any system that involves natural language is the list of words which is collectively referred to as the lexicon. These words would have associated information depending on the purpose of the application. For instance, for automatic translation of documents from one natural language to another, a bi-directional lexicon is essential. Currently, the English-Filipino lexicon contains 23,520 English and 20,540 Filipino word senses with information on the part of speech and co-occurring words, which is based on the dictionary of the Komisyon sa Wikang Filipino. Additional information such as synsetID from Princeton WordNet were integrated into the lexicon [40]. As manually populating the database with the synsetIDs from WordNet is tedious [5], automating the process through the SUMO (Suggested Upper Merged Ontology) as an InterLingual Index (ILI) is now being explored.

An automatic bilingual lexicon extraction from comparable, non-parallel corpora was developed for English and Tagalog as the source and target languages, respectively [52]. Since Tagalog has limited electronic linguistic resources available, we used a limited amount of corpora of 400k and seed lexicon of 9,026 entries in contrast to previous studies of 39M and 16,380, respectively. We combined approaches from previous researches which only concentrated on context extraction, clustering techniques, or usage of part of speech tags for defining the different senses of a word, and ranking has shown improvement to overall F-measure from 7.32% to 10.65% within the range of values from previous studies, despite the use of limited linguistic resources.

Initial work on the manual collection of documents on Philippine languages has been done through the funding from the National Commission for Culture and the Arts considering four major Philippine Languages namely, Tagalog, Cebuano, Ilocano and Hiligaynon with 250,000 words each and the Filipino sign language with 7,000 signs [47]. Computational features include word frequency counts and a concordancer that allows viewing co-occurring words in the corpus.

Aside from possibilities of connecting the Philippine islands and regions through language, we are also aiming at crossing boundaries of time [45; 46]. An unexplored but equally challenging area is the collection of historical documents that will allow research on the development of the Philippine languages through the centuries. An interesting piece of historical information is in Doctrina Christiana, the first ever published work in the country in 1593 which shows the translation of religious material in the local Philippine script, the Alibata, and Spanish. A sample page is shown in Figure 1 (courtesy of the University of Sto. Tomas Library, 2007).

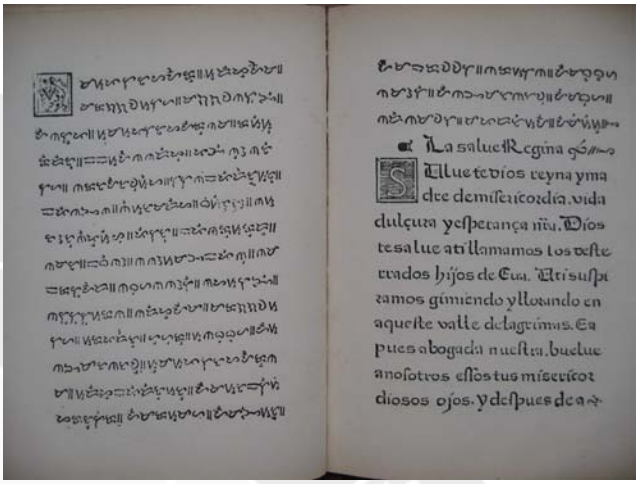


Figure 1. Sample page: doctrina christiana (courtesy of the university of sto. tomas library, 2007)

An automatic retrieval system for documents written in closely-related languages has been developed [20]. Input documents are matched against the n-gram language models of relevant and irrelevant documents. Using common word pruning to differentiate between the closely-related languages, and the odds ratio query generation methods, results show improvements in the precision of the system, using four closely-related Philippine languages.

Although automatic methods can facilitate the building of the language resources needed for processing natural languages, these automatic methods usually employ learning approaches that would require existing language resources as seed or learning data sets.

An online repository of the Philippine corpus [47] provides a venue for linguists or language researchers to upload text documents written in any Philippine language, and to download and analyze corpora on Philippine languages (with URL: http://ccs.dlsu.edu.ph:8086/Palito). Automatic tools for data categorization and corpus annotation are provided by the system. We are refining the mechanics for the levels of users and their corresponding privileges for a manageable monitoring of the corpora. Videos on the Filipino sign language can also be uploaded and downloaded into the system. Uploading of speech recordings will be considered in the near future, to address the need to employ the best technology to document and systematically collect speech recordings of nearly-extinct languages in the country. This online system capitalizes on the opportunity for the corpora to expand faster and wider with the involvement of more people from various parts of the world. This is also to exploit on the reality that many of the Filipinos here and abroad are native speakers of their own local languages or dialects and can largely contribute to the growth of the corpora on Philippine languages.

3. LANGUAGE TOOLS

We have worked on language tools such as morphological processes, part of speech tagging and parsing with experiments on rule-based and example-based approaches.

3.1. Morphological Processes

In general, morphological processes are categorized as morphological analysis or morphological generation. Morphological analyzers (MA) are automated systems that derive the root word of a transformed word, and identify the affixes used and the changes in semantics due to the word transformation. In this way, root words and their derivatives do not have to be stored in the lexicon. On the other hand, morphological generators transform a root word into the surface form given the desired word usage.

Rule-based and example-based approaches for MA and MG provide complementary systems for better computation. In current methods, rule-based MA such as finite-state and unification-based, are predominantly effective only for handling concatenative morphology such as prefixation and suffixation. Because Philippine languages exhibit largely non-concatenative phenomena such as infixation and reduplication, such approaches are found to be lacking, and thus, new approaches are developed to handle both phenomena.

We have explored on the use of optimality theory and two-level morphology rule representation to handle morphological analysis to handle both morphological phenomena [29]. Test results showed 96% accuracy; with a 4% error that is attributed to d-r alteration (e.g. *lakaran*, which is from the root word *lakad* and suffix *-an*, but *d* is changed to *r*). Unfortunately, since all candidates are generated, and erroneous ones are later eliminated through constraints and rules, time efficiency is affected by the exhaustive search performed.

An example-based approach was explored by extending Wicentowski's Word Frame model [11]. In the WordFrame model, the seven-way split re-write rules composed of the canonical prefix/beginning, point-of-prefixation, common prefix substrings, internal vowel change, common suffix substring, point-of-suffixation, and canonical suffix/ending. Infixation, partial and full reduplication as in Tagalog and other Philippine words are improperly modeled in the WordFrame model as point-of-prefixation as in the word (*hin)-intay* which should have been modeled as the word *hintay* with infix *–in-*. Words with an infix within a prefix are also modeled as point-of-prefixation as in the word (*hini-)hintay* which should be represented as infix *–in* in partial reduplicated syllable *hi-*. The non-concatenative Tagalog morphological behaviors such as infixation and reduplication are modeled separately and correctly, in the revised WordFrame model. Using 40,276 Filipino word pairs in automatically leanring re-write rules, a 90% accuracy was obtained when applied to an MA, where some occurrences of reduplication are still represented as point-of-suffixation for various locations of the longest common substring. Analysis of several partial or whole-word reduplications also had some problems. The complexity of a more comprehensive model to handle these would be computationally costly, but it would ensure an increase in accuracy and reduced number of rules.

Although approaches for MA can be extended to handle morphological generation, an additional disambiguation process is necessary to choose the appropriate output from the many various surface forms of words that can be generated from one underlying form.

3.2. Part of Speech Tagging

One of the most useful information in the language corpora are the part of speech tags that are associated with each word in the corpora. Firstly, with the aid of linguists, we have come up with a revised tagset for Tagalog, since a close examination of the existing tagset for languages such as English showed the insufficiency of this tagset to handle certain phenomena in Philippine languages such as lexical markers, ligatures and enclitics. The lexical marker *ay* is used in inverted sentences such as *She is good (Siya ay mabuti)*. Ligatures can take the form of the word *na* or suffixes *-ng* (*-g*), the former is used if the previous noun, pronoun or adjective ends with a consonant (except for *n*), and the latter if the previous word ends with a vowel (or *n*).

Manual tagging of corpora has allowed us to perform automatic experiments on some approaches for tagging for Philippine languages namely MBPOST, PTPOST4.1, TPOST and TagAlog, each one exploring on a particular approach in tagging such as memory-based POS, template-based and rule-based approaches. A study on the performance of these taggers using a POS tagged corpus of 122,287 words showed accuracies of 85, 73, 65 and 61%, respectively [42].

3.3. Language Grammars

Grammar checkers are some of the applications where syntactic specification of languages is necessary. Experiments have been conducted on both rule-based and exampled-based approaches. SpellCheF is a spell checker for Filipino that uses a hybrid approach in detecting and correcting misspelled words in a document [10]. Its approach is composed of dictionary-lookup, n-gram analysis, Soundex and character distance measurements. It is implemented as a plug-in to OpenOffice Writer. Two spelling rules and guidelines, namely, the Komisyon sa Wikang Filipino 2001 Revision of the Alphabet and Guidelines in Spelling the Filipino Language, and the Gabay sa Editing sa Wikang Filipino rulebooks, were incorporated into the system. SpellCheF consists of the lexicon builder, the detector, and the corrector; all of which utilized both manually formulated and automatically learned rules to carry out their respective tasks.

FiSSAn, on the other hand, is a semantics-based grammar checker and implemented as a plug-in to Open Office. Lastly, PanPam, an OpenOffice grammar-checker plugin, is an extension of FiSSAn that also incorporates a dictionary-based spell checker [6; 38].

On the other hand, a grammar rule induction method that has been tested on Filipino is example-based [1]. Constituent structures are automatically induced using unsupervised probabilistic approaches, showing an F1 measure of greater than 69%. To add, experiments revealed that the Filipino language does not follow a strict binary structure as English, but is more right-biased.

Automatic parsing of the Philippine component of the International Corpus of English (ICE-PHI) also used grammar rules that were extracted from training data [27]. Automatic part of speech (POS) tagging was performed using MAKETAG, the tagger that was trained and used on the Great Britain component of the ICE. To correct and verify the tags, initial manual expert tag verification is being conducted on sentences in the corpora where randomly generated 10% of the verbs are found.

3.4. Machine Translation

Various applications have been created to cater to different needs which range from summarizing to question answering and from domains of education to the arts. Described here are some of the language technology applications that have been developed in the country.

The bi-directional English-Filipino machine translation (MT) system is a multi-engine approach (one rule-based method and two example-based methods) for automatic language translation of English and Filipino [44]. Refer to Figure 2 for the Architectural Diagram.

The system accepts as input a sentence or a document in the source language and translates this into the target language. The input text undergoes POS tagging, morphological analysis, and then the actual translation. The output translation undergoes natural language generation including morphological generation. The outputs of the three MT engines will be evaluated by the output modeler to determine the most appropriate among the translation outputs [31]. There are ongoing experiments on the hybridization of the rule-based and the template-based approaches where transfer rules and unification constraints are derived [28].

In identifying the most appropriate translation of a word, syntactic relationships (subject-verb, verb-object, adjective-noun) are used to perform word sense disambiguation based on "word-to-sense" and "sense-to-word" relationship between source words and their translations [22]. Using information from a bilingual dictionary and word similarity measures from WordNet, a target word is selected using statistics from a target language corpus. Test results using English to Tagalog translations showed an overall 64% accuracy for selecting word translation.

3.4.1. Rule-based Machine Translation

The rule-based MT uses lexical functional grammar (LFG) as the formalism to capture the translation rules. An evaluation of how comprehensive and exhaustive the identified grammar is to be considered. Is the system able to capture all possible Filipino sentences? How are all possible sentences to be represented since Filipino exhibits some form of free word order in sentences? The next step is the translation step, that is, the conversion of the computerized representation of the input sentence into the intended target language. After the translation process, the computerized representation of the sentence in the target language is outputted into a sentence form, or called the generation process. Although it has been shown in various studies elsewhere and on various languages that LFG can be used for analysis of sentences, there is still a question of whether it can be used for the generation process. The generation involves the outputting of a sentence from a computer-based representation of the sentence. This is part of the work that the group intends to address.

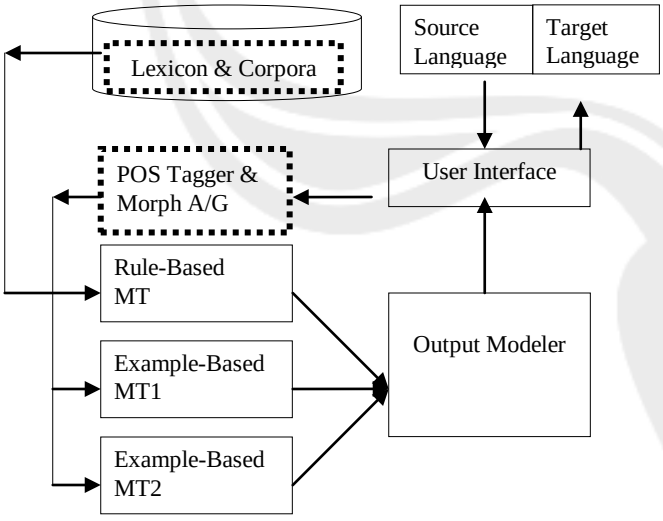


Figure 2. The architecture of the hybrid english-filipino machine translation system

The major advantage of the rule-based MT over other approaches is that it can produce high quality translation for sentence patterns that were accurately captured by the rules of the MT engine; but unfortunately, it cannot provide good translations to any sentence that go beyond what the rules have considered.

3.4.2. Corpus-based Machine Translation

The corpus-based MT system automatically learns how translation is done through examples found in a corpus of translated documents. The system can incrementally learn when new translated documents are added into the knowledge-base, thus, any changes to the language can also be accommodated through the updates on the example translations. This means it can handle translation of documents from various domains [2].

The principle of garbage-in-garbage-out applies here; if the example translations are faulty, the learned rules will also be faulty. That is why, although human linguists do not have to specify and come up with the translation rules, the linguist will have to first verify the translated documents and consequently, the learned rules, for accuracy.

It is not only the quality of the collection of translations that affects the overall performance of the system, but also the quantity. The collection of translations has to be comprehensive so that the translation system produced will be able to translate as much types of sentences as possible. The challenge here is coming up with a quantity of examples that is sufficient for accurate translation of documents.

With more data, a new problem arises when the knowledge-base grows so large that access to it and search for applicable rules during translation requires tremendous amount of access time and to an extreme, becomes difficult. Exponential growth of the knowledge-base may also happen due to the free word order nature of Filipino sentence construction, such that one English sentence can be translated to several Filipino sentences. When all these combinations are part of the translation examples, a translation rule will be learned and extracted by the system for each combination, thus, causing growth of the knowledge-base. Thus, algorithms that perform generalization of rules are considered to remove specificity of translation rules extracted and thus, reduce the size of the rule knowledge-base.

4. AUDIO AND VIDEO PROCESSING

Aside from textual language representation and documentation, we have also explored on audio and video forms and their corresponding application tools.

The Filipino Speech Corpus (FSC) is a compilation of read texts and spontaneous speech recorded from 100 speakers [34]. The read texts consist of paragraphs, sentences, words, syllables and phonemes that were designed to elicit the phones and prosodic cues characterizing Filipino speech. The spontaneous speech recordings are on common topics that the speaker can freely talk about.

Another work on corpus compilation in the country is the Philippine component of the International Corpus of English (ICE-PHI) [19] where 278 spoken files (with 2,000 words each) were transcribed and documented. Indigenous words, phrases and sentences in various Philippine languages can be found in the utterances, and studies on code-switching can be done on these corpora.

Speech processing studies that use the FSC include automatic speech recognition [8; 15; 18; 33; 48; 51] and text-to-speech [8; 14; 15; 24; 53]. Other Filipino speech processing applications that were developed without the use of FSC include PinoyTalk [9], which uses a rule-based Filipino syllabication model applied on an Indonesian language synthesizer, and Tagapagsalita [4], which uses the voice model of a Filipino male and a Filipino female to count from 1 to 100. Speaker identification and verification applications [30; 37] were also developed using a small corpus of 10 speakers each with 5 recordings of their individual passwords.

Another branch of speech research that is recently gaining popularity in the Philippines is detecting emotions from speech. Ebarvia et. al. [23] developed a system that automatically recognizes emotions such as anger, boredom, happiness and satisfaction using a call center database. Chua et. al. [13], on the other hand, recognizes emotions such as happiness, sadness, anger, fear, surprise, disgust, and neutral, using a corpus of 10,500 acted-emotion Filipino speech recordings. This local trend in speech research is stimulated by the large number of contact centers in the country requiring speech analytics applications, and by improving human-machine interactions through empathic computing.

The documentation of the many languages of the country is an obviously enormous task, and it has been recommended [50] that documentation through speech recordings is the most efficient way to document even those languages that are near extinction. Speech processing applications on these corpora are to be explored. Applications such as bilingual/multilingual translators, Filipino speech-to-text and text-to-speech systems for mobile and low-cost devices, speech training software, and dialogue analysis for data mining are just some of the interesting research topics that researchers of the Filipino language can pursue.

The Philippine language corpus engagement includes the Filipino Sign Language. The signs and discourse are recorded in videos, which are edited, glossed and transcribed. Video editing merely cuts the video for final rendering, glossing allows association of sign to particular words, and transcription allows viewing of textual equivalents of the signed videos [47].

Work has been done on sign language number recognition [49] using color-coded gloves for feature extraction using digital signal processing. The feature vectors were calculated based on the position of the dominant-hand's thumb. The system learned through a database of numbers from 1 to 1000, and tested by the automatic recognition of Filipino sign language numbers and conversion into text. Over-all accuracy of number recognition is 85%.

Another proposed work is the recognition of non-manual signals focusing on the various part of the face; in particular, initially, the mouth is to be considered. The automatic interpretation of the signs can be disambiguated using the interpretation of the non-manual signals.

5. HLT ON ENGLISH: PHILIPPINE CONTEXT

As mentioned, past colonial influences are manifested in language usage in Philippine society by the use of non-indigenous languages (such as English, Spanish and Chinese). Some linguistic studies are focused on understanding the linguistic phenomena in the use of these foreign languages in the Philippine context. For instance, English is still being used as the medium of instruction in our schools, although there are new advocacies to using the multi-lingual mother tongue language education in our schools in the early years of schooling. Currently, studies have been on English monolingual human language technology (HLT), and future directions can include the study of multilingual HLT.

5.1. Natural Language Generation

Most of the work in the country on natural language generation has been on the English language and are rule-based in approach some of which are in the areas of text summarization, text simplification, story generation and multiple choice question generation, intelligent tutoring systems, and question answering systems.

SUMMER RXT automatically summarizes a document given the desired percentage of reduction [21]. Chunks of words in sentences are annotated or tagged as nucleus and satellites and some relationships based on Rhetorical Structure Theory are automatically established. Conceptually, the summaries are generated from lifting the nucleus and leaving out the satellites. Keywords and key phrases are also considered during summarization on top of nucleus extraction. Moreover, connector words are added to smoothen transitions from one sentence to another. Thus, the summarized text maintains coherence without having to resort to copying whole sentences from the original text. To add, the removal of an arbitrary amount of source material has the potential of losing essential information. Evaluation against existing commercially available software has shown that the output of SUMMER RXT is comparable to these systems. Unfortunately, the domain of the training and test data has been limited to one particular author and in one particular domain. Experiments on this approach for a wider range of authors and styles and their corresponding domains are yet to be performed.

SimText is a text simplification system that accepts as input a medical document and transforms complex sentences into a set of equivalent simpler sentences with the goal of making the resulting text easier to read by some target group [17]. The simplification includes the use of easier to understand terminologies and shorter

sentence constructs considering the specified reading level of the intended target users. The text simplification process identifies components of sentence that may be separated out, and transforms each of these into free-standing simpler sentences. Some nuances of meaning from the original text may be lost in the simplification process, since sentence-level syntactic restructuring can possibly alter the meaning of the sentence.

Picture Books generates stories for children from an input picture containing the background and a set of character and object stickers [36]. The child chooses the stickers and the system associates these to a theme and a (manually created) ontology which are then used to generate a fable-type of story. Figure 3 shows a sample screen shot, where the left side of the screen contains the picture (background and stickers selected by the child) while the right side contains part of the generated story.

On the other hand, MesCH is a software that accepts children's stories and automatically generates multiple choice questions to test the child's reading comprehension [25]. The program rephrases parts of the story into 4W questions (who, what, when, where), sequence questions (which came first), and vocabulary questions. To illustrate, from a sentence in a story "Slimy tadpoles came out from the eggs", the system will generate the following possible stems:

1 What came out from the eggs?
2 Where did the slimy tadpoles come out?
3 In the sentence, "Slimy tadpoles came out from the eggs," what does the verb "came out" mean?

The system considers principles in instructional assessment such as the formulation of 4W questions and the construction of distractors through the use of entries in WordNet that relate with the correct answer.

Popsicle is an intelligent tutoring system that identifies and corrects language errors committed by Filipino students while they are learning the English language [35]. The software initially assesses the English grammar proficiency of the learner based on an input essay document that was composed by the user, identifies the grammatical errors in the document, provides feedback and suggestions in natural language, and generates grammar lessons that are tailor fit to the individual needs of the learner. The learner is given opportunities to correct and learn from his mistakes. The software maintains a user model that tracks an individual learner's English grammar proficiency, his position and path toward acquiring English, the dialogue history containing the text generated by the system during the current tutorial session, the evaluation scores for each of the teaching strategies employed, and a concise log of explanations attempted by the system over the learning period of the user.

HelloPol is a question-answering system that converses with the user in English within the Philippine political domain [3]. The system has been fed with political news articles, and information extraction has been integrated into the system to automatically extract relevant information from the articles into a more structured type of representation (or simply, a database) for use in the question-answering system. The user may ask factoid questions (who, what, when, where) and the program answers these by referring to the database of information. It is also an adaptive question-answering system in that it considers in its responses the user's topic preference during the course of the dialogue.

The area of question-answering is moving towards research on other types of questions apart from factoid, definition, and list questions. Questions that involve evaluation and comparison, where the answer cannot be directly lifted from source text, are some of these new types of questions [7]. Evaluative refers to the consideration of at least one property or criteria over one or more entities and the computation of the associated values. Comparative refers to the evaluation of objects depending on one or more criteria and classifying those objects depending on the returned values. Included in comparative is the identification of the extreme, i.e., the superlatives, the topmost objects. In such cases, the focus of the questions is on the properties at stake in the evaluation, leading to the comparison. Thus, comparative and evaluative QA involves answering questions that require various forms of inference related to evaluation before an answer can be given. Since evaluation is necessary, the answer is not lifted from source text, as in the case of answering factoid, definition, or list questions. Instead, natural language answers will have to be constructed from the results of numeric and non-numeric evaluations of the criteria [39].

Human conversation frequently involves the use of creative text such as puns to make the interaction fun and engaging. For dialogue systems to achieve the same effect, computers must be trained to understand when a pun has been delivered and to find opportunities to generate puns as a means of conveying information. TPEG [36] has been developed as our first step towards this goal.

Utilizing ConceptNet [41] and Unisyn [26], TPEG identifies and extracts word relationships from an input corpus of human puns, which it then uses to generate its own puns. Extracted word relationships include phonetic similarity, synonyms, and semantic relationships (such as part-of, location-of, property-of). Listing 3 shows 2 pairs of human source puns and TPEG-generated puns.

```
Source Pun 1:
   What do you call a beloved mammal?
   A dear deer. [60]
TPEG Pun 1:
   What do you call an overall absence?
   A whole hole.

Source Pun 2:
   How is a window like a headache?
   They are both panes. [60]
TPEG Pun 2:
   How is a trunk like a garbage?
   They are both waists.
```

Listing 3. Sample human and TPEG generated puns

5.2. Information Extraction

LegalTRUTHS performs automatic extraction of structured data from unstructured data; that is, from long textual documents (from the Supreme Court of the Philippines) to databases [12]. It aims to minimize the user's need of going through countless number and infinitely long legal documents and court decisions to extract key

information about the case at hand. Based on the sample documents, a template for the database was developed through consultations with lawyers. The process follows the traditional approach wherein preprocessing of the input text is performed which includes text segmentation into different regions, detection of sentence boundaries, part of speech tagging and named entity recognition. Then text recognition is performed by applying the corresponding rules as needed to fill up the database. These include detection of noun and verb groups as a whole entity, normalization of the output, filtering of irrelevant information, co-reference resolution and extraction of the basic fields in the proposed template. The system also has an automatic evaluation module that uses longest common subsequence and the metrics precision, recall and f-measure to check the system's correctness. As a front-end application, the system also provides keyword search from the extracted fields. The matching entries provide links to the actual documents. Figure 3 shows a sample screen shot of the relevant information extracted into table form. Overall results show precision at 91%, recall at 99%, and F-measure at 95%.

6. FUTURE DIRECTIONS

Much is to be explored in this area of research that interleaves diverse disciplines among technology-based areas (such as NLP, digital signal processing, multi-media applications, and machine learning) and other fields of study (such as language, history, psychology, and education), and cuts across different regions and countries, and even time frames. It is multi-modal that considers various forms of data from textual, audio, video and other forms of information. Thus, much is yet to be accomplished, and experts with diverse backgrounds in these various related fields will bring this area of research to a new and better dimension.

6. ACKNOWLEDGEMENTS

The authors would like to acknowledge the support of our consistent academic and government partners, to name a few: the Commission on Higher Education (CHED), CHED-Zonal Research Center (CHED-ZRC), the Philippine Council for Advanced Science and Technology Research and Development, Department of Science and Technology (PCASTRD/DOST), Komisyon sa Wikang Filipino (KWF), and the National Commission for Culture and the Arts (NCCA).

7. REFERENCES

[1] Alcantara, D. and A. Borra. 2008. Constituent Structure for Filipino: Induction through Probabilistic Approaches. *Proceedings of the 22nd Pacific Asia Conference on Language, Information and Computation*. 113-122.

[2] Alcantara, D., Hong, B., Perez, A. and Tan, L. 2006. "Rule Extraction Applied in Language Translation – R.E.A.L. Translation". Undergraduate Thesis, De la Salle University.

[3] Alimario, P. M., Cabrera, A., Ching, E., Sia, E. J. and Tan, M. W. 2003. HelloPol: An Adaptive Political Conversationalist. *Proceedings of the 1st National Natural Language Processing Research Symposium* (2003).

[4] Aralar, K., Coloso, P. M., Moneda, J., Ilao, J., and Cu, J. 2008. "Tagapagsalita: A Text-to-Speech System for Counting in Filipino", ADD-3, Thailand.

[5] Borra, A, Pease, A., Roxas, R., and Dita, S. 2010. Introducing Filipino WordNet. 5th Global WordNet Association conference, January 31 to February 4, 2010, India Institute of Technology, Bombay, India.

[6] Borra, A., Ang, M., Chan, P. J., Cagalingan S., and Tan. R. 2007. FiSSan: Filipino Sentence Syntax and Semantic Analyzer. *Proceedings of the 7th Philippine Computing Science Congress.* 74-78 (February 2007).

[7] Burger, J., et al. Issues, Tasks and Program Structures to Roadmap Research in Question & Answering. Available in: wwwnlpir.nist.gov/projects/duc/papers/qa.Roadmappaperv2.doc. (2009).

[8] Cayaban, C., Climaco, J., Espina, E., and Guevara, R.C.L. 2001. "A Low Bit Rate Filipino Speech Coder Using Hidden Markov Model-Based Speech Recognition/Speaker Independent Synthesis Techniques", in Proc. 2nd National ECE Conference.

[9] Casas, D, Rivera, S., Tan, G., and Villamil, G. 2004. PinoyTalk: A Filipino Based Text-to-Speech Synthesizer. Undergraduate Thesis. De La Salle University (April 2004).

[10] Cheng, C., Alberto, C. P., Chan, I. A., and Querol, V. J. 2007. SpellChef: Spelling Checker and Corrector for Filipino. *Journal of Research in Science, Computing and Engineering*. 4(3), 75-82 (December 2007).

[11] Cheng, C., and See, S. 2006. The Revised Wordframe Model for Filipino Language. *Journal of Research in Science, Computing and Engineering*. 3(2), 17-23 (August 2006).

[12] Cheng, T. T., Cua, J. L., Tan, M. D., and Yao, K. G. 2008. Legal TRUTHS: Turning Unstructured Text Helpful Structure for Legal Documents. Undergraduate Thesis. De La Salle University (September 2008).

[13] Chua, J., De Guia, O., Li, C., and Rojas, F. 2009. "Emotion Recognition in Filipino Speech", Undergraduate thesis, De La Salle University, 2009.

[14] Co, M., R., and Guevara, C. L. 2003. "Prosody Modification In Filipino Speech Synthesis Using Dynamic Time Warping," IEEE TENCON in Bangalore, India.

[15] Corpus, M., Liampo, J., Co, M., and Guevara, R. C. L. 2001. "Development of a Filipino TTS System using Concatenative Speech Synthesis," in Proc. 2nd National ECE Conference.

[16] Cua, J., Manurung, R., Ong, E. and Pease, A. 2010. Representing Story Plans in SUMO. Proceedings of the NAACL Human Language Technology 2010 Workshop on Computational Approaches to Linguistic Creativity, Los Angeles, USA.

[17] Damay, J. J., Lojico, G. J., Lu, K. A., Tarantan, D., and Ong, E. 2006. SIMTEXT: Text Simplification of Medical Literature. *Proceedings of the 3rd National Natural Language Processing Research Symposium* (2006).

[18] dela Vega, E., Co, M., and Guevara, R. C. L. 2002. "Language Model for Predicting Parts of Speech of Filipino Sentences," in Proc. 3rd National ECE Conference, 2002.

[19] Department of English and Applied Linguistics, De La Salle University. "International Corpus of English, The Philippine Corpus," April 2004.

[20] Dimalen, D. M. and Roxas, R. 2007. AutoCor: A Query-Based Automatic Acquisition of Corpora of Closely-Related Languages. *Proceedings of the 21st Pacific Asia Conference on Language, Information and Computation*. 146-154 (November 2007).

[21] Diola, A. M., Lopez, J. T., Torralba, P., So, S., and Borra, A. 2004. Automatic Text Summarization. *Proceedings of the 2nd National Natural Language Processing Research Symposium* (2004).

[22] Domingo, E. and Roxas, R. 2006. Utilizing Clues in Syntactic Relationships for Automatic Target Word Sense Disambiguation. *Journal of Research for Science, Computing and Engineering*. 3(3), 18-24 (December 2006).

[23] Ebarvia, E., Bayona, M., de Leon, F., Lopez, M., Guevara, R., Calingacion, B., Naval, Jr., P. 2008. "Determination of Prosodic Feature Set for Emotion Recognition in Call Center Speech", Proceedings of the 5th National Natural Language Processing Research Symposium (NNLPRS). 65-71. 2008.

[24] Espina, E., Tan, E., and Guevara, R. 2002. "Real-time Implementation of a low bit rate Filipino speech codec using hidden markov model-based speech recognition/synthesis," in Proc. 3rd National ECE Conference, 2002.

[25] Fajardo, K., Di, S., Novenario, K., and Yu, C. 2008. Mesch: Measurement System for Children's Reasing Comprehension. Undergraduate Thesis. De La Salle University (September 2008).

[26] Fitt, S. Unisyn Lexicon Release. Available: http://www.cstr.ed.ac.uk/projects/unisyn/ (2002).

[27] Flores, D. and Roxas, R. 2008. Automatic Tools for the Analysis of the Philippine component of the International Corpus of English. *Linguistic Society of the Philippines Annual Meeting and Convention* (2008).

[28] Fontanilla, G., and Roxas, R. 2008. A Hybrid Filipino-English Machine Translation System. *DLSU Science and Technology Congress* (July 2008).

[29] Fortes-Galvan, F. C. and Roxas, R. 2007. Morphological Analysis for Concatenative and Non-concatenative Phenomena. *Proceedings of the Asian Applied NLP Conference* (March 2007).

[30] Go, G., Manza, L. O., Realeza, M. A., and Ting, R. 2001. "Voice Activated Lock", Undergraduate thesis, De La Salle University, 2001.

[31] Go, K. and See, S. 2008. Incorporation of WordNet Features to N-Gram Features in a Language Modeller. *Proceedings of the 22nd Pacific Asia Conference on Language, Information and Computation*, 179-188 (November 2008).

[32] Gordon, R. G., Jr. (Ed.). 2005. *Ethnologue: Languages of the World*, Fifteenth edition. Dallas,Texas: SIL International. Online version: www.ethnologue.com Online (2005).

[33] Guevara, R. C. L. 2003. "Formulating a Filipino-English Phoneme Set Using a Multi-Layer Perception and a Start Synchronous Decoder", Proc. 4th National ECE Conference, 2003.

[34] Guevara, R. C., L. M. Co, E. Tan, I. D. Garcia, E. Espina, R. Ensomo, and R. Sagum, "Development of a Filipino Speech Corpus," in Proc. 3rd National ECE Conference, 2002.

[35] Gurrea, A. M., Liu, A., Ngo Vincent, D., Que, J., and Ong, E. 2006. Recognizing Syntactic Errors in Written Philippine English. *Proceedings of the 3rd National Natural Language Processing Research Symposium* (2006).

[36] Hong, B. 2008. Template-based Pun Extractor and Generator. Graduate Thesis. De La Salle University (March 2008).

[37] Jacinto, J., Nario, J. R., See, A. L., and Umali, J. M. 2002. "Wavelet-Based Speaker Verification System", Undergraduate thesis, De La Salle University, 2002.

[38] Jasa, M. A., Palisoc, M. J., and Villa, J. M. 2007. Panuring Panitikan (PanPam): A Sentence Syntax and Semantics-based Grammar Checker for Filipino. Undergraduate Thesis. De La Salle University (September 2007).

[39] Lim, N. R., Saint-Dizier, P., and Roxas, R. 2009. Some Challenges in the Design of Comparative and Evaluative Question Answering Systems. *Proceedings of the 2009 Workshop on Knowledge and Reasoning for Answering Questions*, (2009).

[40] Lim, N. R., Lat, J. O., Ng, S. T., Sze, K., and Yu, G. D. 2007. Lexicon for an English-Filipino Machine Translation System. *Proceedings of the 4th National Natural Language Processing Research Symposium* (2007).

[41] Liu, H. and Singh, P. 2004. ConceptNet — A Practical Commonsense Reasoning Tool-Kit. *BT Technology Journal*, 22(4): 211-226. Netherlands: Springer (2004).

[42] Miguel, D. and Roxas, R. 2007. Comparative Evaluation of Tagalog Part of Speech Taggers. *Proceedings of the 4th National Natural Language Processing Research Symposium* (2007).

[43] Roxas, R., Alcantara, D., and Borlongan, A. 2010. Language Documentation and Applications in the Philippines: Implications for Mother Tongue-Based Multilingual Education. 1st Philippine Conference Workshop on Mother Tongue Based Multilingual Education. February 18-20, 2010, Cagayan de Oro City, Philippines.

[44] Roxas, R. E., Borra, A., Ko, C., Lim, N. R., Ong, E., and Tan, M. W. 2008. Building Language Resources for a Multi-Engine Machine Translation System. *Language Resources and Evaluation*. Springer, Netherlands. 42:183-195 (2008).

[45] Roxas, R. 2007. e-Wika: Philippine Connectivity through Languages. *Proceedings of the 4th National Natural Language Processing Research Symposium* (2007).

[46] Roxas, R. 2007. Towards Building the Philippine Corpus. *Consultative Workshop on Building the Philippine Corpus* (November 2007).

[47] Roxas, R., Inventado, P., Asenjo, G., Corpus, M., Dita, S., Sison-Buban, R., and Taylan, D. 2009. Online Corpora of Philippine Languages. *2nd DLSU Arts Congress: Arts and Environment* (February 2009).

[48] Sagum, R. G., Ensomo, R. A., Tan, E. M., and Guevara, R.C. L. 2003. "Phoneme Alignment of Filipino Speech Corpus", TENCON 2003. Conference on Convergent Technologies for Asia-Pacific Region, Volume 3, 15-17 Oct. 2003 Page(s): 964 - 968 Vol.3.

[49] Sandjaja, I., and Marcos, N. 2009. Sign Language Number Recognition. *NCM 2009 (International Conference on Networked Computing, Advanced Information Management and Digital Content and Multimedia Technologies), 5th International Joint Conference on INC, IMS and IDC*, August 25-27, 2009.

[50] Simons. G. 2008. "Toward a Global Infrastructure for the Sustainability of Language Resources." Plenary Talk, Proceedings of the Pacific Asia Conference on Language, Information and Computation 22, 2008.

[51] Tantan, S. M. A., Tan, E., and Guevara, R. C. L. 2003. "Speech/Non-Speech Detection of the Filipino Speech Corpus", Proc. 4th National ECE Conference.

[52] Tiu, E. P., and Roxas, R. 2008. Automatic Bilingual Lexicon Extraction for a Minority Target Language, *Proceedings of*

the 22nd Pacific Asia Conference on Language, Information and Computation. Best Paper Awardee by PACLIC Steering Committee. 368-376 (November 2008).

[53] Tupas, L., Co, M., and Guevara, R. C. L. 2002. Concatenative Text-to-Speech Synthesis of Two-Syllable Filipino Words. *Proceedings of the 3rd National ECE Conference* (2002).

Hybrid-Multidimensional Fuzzy Association Rules from a Normalized Database[1]

Rolly Intan
Informatics Engineering Department
Petra Christian University, Surabaya, Indonesia
rintan@peter.petra.ac.id

ABSTRACT

Mining association rules is one of the important tasks in the process of data mining application. In general, the input as used in the process of generating rules is taken from a certain data table by which all the corresponding values of every domain data have correlations one to each others as given in the data table. A problem arises when we need to generate the rules expressing the relationship between two or more domains that belong to several different tables in a normalized database. To overcome the problem, before generating rules it is necessary to join the participant tables into a general table by a process called Denormalization..

This paper shows a process of mining Multidimensional Fuzzy Association Rules from a normalized database. The process consists of two sub-process, namely sub-process of join tables (Denormalization) and sub-process of mining fuzzy rules. Some parts of mining the fuzzy association rules as discussed in our previous papers [3,4,5,6,7]are extended to generate hybrid-multidimensional fuzzy association rules.

1. INTRODUCTION

Association rule finds interesting association or correlation relationship among a large data set of items [1,10]. The discovery of interesting association rules can help in decision making process. Association rule mining that implies a single predicate is referred as a single dimensional or intradimension association rule since it contains a single distinct predicate with multiple occurrences (the predicate occurs more than once within the rule). The terminology of single dimensional or intradimension association rule is used in multidimensional database by assuming each distinct predicate in the rule as a single dimension [1].

Here, the method of *market basket analysis* can be extended and used for analyzing any context of database. For instance, database of medical track record patients is analyzed for finding association (correlation) among diseases taken from the data of complicated several diseases suffered by patients in a certain time. For example, it might be discovered a Boolean association rule "Bronchitis $\Rightarrow$ Lung Cancer" representing relation between "Bronchitis" and "Lung Cancer" which can also be written as a single dimensional association rule as follows:

Rule-1

$$Dis(X,\text{"Bronchitis"}) \Rightarrow Dis(X,\text{"Lung Cancer"}),$$

where D is a given predicate and X is a variable representing patient who have a kind of disease (i.e. "Bronchitis" and "Lung Cancer"). In general, "Lung Cancer" and "Bronchitis" are two different data that are taken from a certain data attribute, called *items*. In general, *Apriori* [1,10] is used an influential algorithm for mining frequent itemsets for mining Boolean (single dimensional) association rules.

Additional related information regarding the identity of patients, such as *age, occupation, sex, address, blood type*, etc., may also have a correlation to the illness of patients. Considering each data attribute as a predicate, it can therefore be interesting to mine association rules containing *multiple* predicates, such as:

Rule-2:

$$Age(X,\text{"60"}) \wedge Smk(X,\text{"yes"}) \Rightarrow Dis(X,\text{"Lung Cancer"}),$$

where there are three predicates, namely *Age*, *Smk* (*smoking)* and *Dis (disease)*. Association rules that involve two or more dimensions or predicates can be referred to as *multidimensional association rules*. Multidimensional association rules with no repeated predicate as given by Rule-2, are called *interdimension association rules* [1]. It may be interesting to mine multidimensional association rules with repeated predicates. These rules are called *hybrid-multidimensional association rules*, e.g.:

Rule-3:

$$Age(X,\text{"60"}) \wedge Smk(X,\text{"yes"}) \wedge Dis(X,\text{"Bronchitis"}) \Rightarrow Dis(X,\text{"Lung Cancer"}),$$

To provide a more meaningful association rule, it is necessary to utilize *fuzzy sets* over a given database attribute called *fuzzy association rule* as discussed in [4,5]. Formally, given a crisp domain D, any arbitrary fuzzy set (say, fuzzy set A) is defined by a membership function of the form [2,9]:

$$A: D \rightarrow [0,1]. \qquad (1)$$

A fuzzy set may be represented by a meaningful fuzzy label. For example, "*young*", "*middle-aged*" and "*old*" are fuzzy sets over *age* that is defined on the interval [0, 100] as arbitrarily given by[2]:

[1] Revised and Extended version of the paper presented at ICHIT 2008[7]

$$young(x) = \begin{cases} 1 & , x \le 20 \\ (35-x)/15 & , 20 < x < 35 \\ 0 & , x \ge 35 \end{cases}$$

$$middle\_aged(x) = \begin{cases} 0 & , x \le 20 \text{ or } x \ge 60 \\ (x-20)/15 & , 20 < x < 35 \\ (60-x)/15 & , 45 < x < 60 \\ 1 & , 35 \le x \le 45 \end{cases}.$$

$$old(x) = \begin{cases} 0 & , x \le 45 \\ (x-45)/15 & , 45 < x < 60 \\ 1 & , x \ge 60 \end{cases}$$

Using the previous definition of fuzzy sets on *age*, an example of hybrid-multidimensional fuzzy association rule relation among the predicates *Age*, *Smk* and *Dis* in Rule-3 may then be represented by:

Rule-4

$$Age(X, \text{"young"}) \wedge Smk(X, \text{"yes"}) \wedge Dis(X, \text{"Bronchitis"}) \Rightarrow Dis(X, \text{"Lung Cancer"}),$$

To generate hybrid-multidimensional association rules implying fuzzy value such as given by Rule-4 from a normalized database that consists of several tables, this paper discussed two sequential processes as shown in Figure 1.

First is the process of joining tables known as Denormalization of Database. Second is the process of generating (mining) fuzzy association rules. The process of denormalization can be provided based on the relation of tables as presented in Entity Relationship Diagram (ERD) of the relational database.

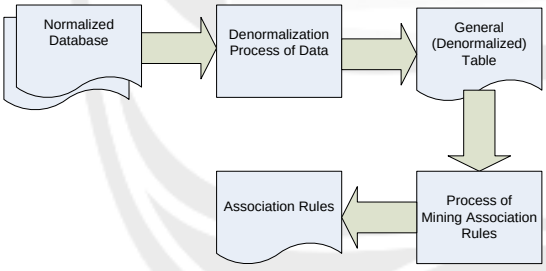


Figure 1. Process of mining association rules

For two tables that have no direct relation in ERD, they can still be joined by others transition tables (in ERD) using the transitive join process. Other solution is that we can define or create a relation function or a relation table that corresponds two distinct domains of the tables. Here, a metadata can be constructed as a data dictionary to express the relationship of tables. Result of denormalization data process is a single general (denormalized) table. The table is used as a source data for the process of mining association rules. Some parts of mining fuzzy rules has been discussed in [4,5,6,7] that introduced some formulations for calculating support and confidence factors. This paper extends the concepts to be applied in mining hybrid-dimensional fuzzy association rules and also introduces a formula to calculate correlation factor as also usually used in evaluating interestingness of the rules.

The structure of the paper is the following. In Section 2, basic definition and formulation of some measures, support, correlation and confidence rule as used for determining interestingness of association rules are briefly recalled. Section 3 as a main contribution of this paper is devoted to propose data preparation for the further process of generation rules. Here, we will discuss a process of join table from a normalized database. Section 4 discusses a concept for mining multidimensional fuzzy association rules. Section 5 demonstrated the concept in an illustrative example. Finally a conclusion is given in Section 6.

2. SUPPORT, CONFIDENCE AND CORRELATION

Association rules are kind of patterns representing correlation of attribute-value (items) in a given set of data provided by a process of data mining system. Generally, association rule is a conditional statement (such kind of *if-then rule*). More formally [1], association rules are the form $A \Rightarrow B$, that is,

$a_1 \wedge \cdots \wedge a_m \Rightarrow b_1 \wedge \cdots \wedge b_n$, where a_i (for $i \in \{1,\ldots,m\}$) and b_j (for $j \in \{1,\ldots,n\}$) are two items (attribute-value). The association rule $A \Rightarrow B$ is interpreted as "*database tuples that satisfy the conditions in A are also likely to satisfy the conditions in B*". $A = \{a_1, \cdots, a_m\}$ and $B = \{b_1, \cdots, b_n\}$ are two distinct itemsets. Performance or interestingness of an association rule is generally determined by three factors, namely *confidence*, *support* and *correlation* factors. Confidence is a measure of certainty to assess the validity of the rule. Given a set of relevant data tuples (or transactions in a relational database) the confidence of "$A \Rightarrow B$" is defined by:

$$\text{conf}(A \Rightarrow B) = \frac{\#tuples(A \text{ and } B)}{\#tuples(A)}, \qquad (2)$$

where #*tuples*(*A* and *B*) means the number of tuples containing *A* and *B*.

For example, a confidence 80% for the Association Rule (for example Rule-1) means that 80% of all patients who infected bronchitis are likely to be also infected lung cancer. The support of an association rule refers to the percentage of relevant data tuples (or transactions) for which the pattern of the rule is true. For the association rule "$A \Rightarrow B$" where *A* and *B* are the sets of items, support of the rule can be defined by

$$\text{supp}(A \Rightarrow B) = \text{supp}(A \cup B) = \frac{\#tuples(A \text{ and } B)}{\#tuples(all\_data)}, \qquad (3)$$

where #*tuples*(*all_data*) is the number of all tuples in the relevant data tuples (or transactions).

For example, a support 30% for the association rule (e.g., Rule-1) means that 30% of all patients in the all data medical records are infected both bronchitis and lung cancer. From (3), it can be

followed $\text{supp}(A \Rightarrow B) = \text{supp}(B \Rightarrow A)$. Also, (2) can be calculated by

$$\text{conf}(A \Rightarrow B) = \frac{\text{supp}(A \cup B)}{\text{supp}(A)}, \qquad (4)$$

Correlation factor is another kind of measures to evaluate correlation between A and B. Simply, correlation factor can be calculated by:

$$\text{corr}(A \Rightarrow B) = \text{corr}(B \Rightarrow A) = \frac{\text{supp}(A \cup B)}{\text{supp}(A) \times \text{supp}(B)}, \qquad (5)$$

Itemset A and B are dependent (positively correlated) iff $\text{corr}(A \Rightarrow B) > 1$. If the correlation is equal to 1, then *A* and *B* are independent (no correlation). Otherwise, A and B are negatively correlated if the resulting value of correlation is less than 1.

A data mining system has the potential to generate a huge number of rules in which not all of the rules are interesting. Here, there are several objective measures of rule interestingness. Three of them are measure of rule support, measure of rule confidence and measure of correlation. In general, each interestingness measure is associated with a threshold, which may be controlled by the user. For example, rules that do not satisfy a confidence threshold (*minimum confidence*) of, say 50% can be considered uninteresting. Rules below the threshold (*minimum support* as well as *minimum confidence*) likely reflect noise, exceptions, or minority cases and are probably of less value. We may only consider all rules that have positive correlation between its itemsets.

3. DENORMALIZATION DATA

In general, the process of mining data for discovering association rules has to be started from a single table (relation) as a source of data representing relation among item data. Formally, a relational data table [13] *R* consists of a set of tuples, where t_i represents the *i*-th tuple and if there are *n* domain attributes *D*, then $t_i = (d_{i1}, d_{i2}, \cdots, d_{in})$. Here, d_{ij} is an atomic value of tuple t_i with the restriction to the domain D_j, where $d_{ij} \in D_j$. Formally, a relational data table *R* is defined as a subset of the set of cross product $D_1 \times D_2 \times \cdots \times D_n$, where $D = \{D_1, D_2, \cdots, D_n\}$. Tuple *t* (with respect to *R*) is an element of *R*. In general, *R* can be shown in Table 1.

Table 1. A schema of relational data table

Tuples	D_1	D_2	$\cdots$	D_n
t_1	d_{11}	d_{12}	$\cdots$	d_{1n}
t_2	d_{21}	d_{22}	$\cdots$	d_{2n}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
t_r	d_{r1}	d_{r2}	$\cdots$	d_{rn}

A normalized database is assumed as a result of a process of normalization data in a certain context of data. The database may consist of several relational data tables in which they have relation one to each others. Their relation may be represented by Entities Relationship Diagram (ERD). Hence, suppose we need to process some domains (columns) data that are parts of different relational data tables, all of the involved tables have to be combined (joined) together providing a *general data table*. Since the process of joining tables is an opposite process of normalization data by which the result of general data table is not a normalized table, simply the process is called *Denormalization*, and the general table is then called *denormalized table*. In the process of denormalization, it is not necessary that all domains (fields) of the all combined tables have to be included in the targeting table. Instead, the targeting denormalized table only consists of interesting domains data that are needed in the process of mining rules. The process of denormalization can be performed based on two kinds of data relation as follows.

3.1. Metadata of the Normalized Database

Information of relational tables can be stored in a metadata. Simply, a metadata can be stored and represented by a table. Metadata can be constructed using the information of relational data as given in Entity Relationship Diagram (ERD). For instance, given a symbolic ERD physical design is arbitrarily shown in Figure 2. From the example, it is clearly seen that there are four tables: **A**, **B**, **C** and **D**. Here, all tables are assumed to be independent for they have their own primary keys. Cardinality of relationship between Table **A** and **C** is supposed to be one to many relationships. It is similar to relationship between Table **A** and **B** as well as Table **B** and **D**.

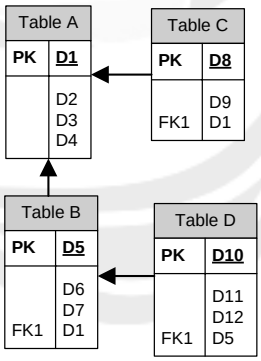


Figure 2. Example of ERD physical design

Table **A** consists of four domains/fields, D1, D2, D3 and D4; Table **B** also consists of four domains/fields, D1, D5, D6 and D7; Table **C** consists of three domains/fields, D1, D8 and D9; Table **D** consists of four domains/fields, D10, D11, D12 and D5. Therefore, there are totally 12 domains data as given by D={D1, D2, D3, …, D11, D12}. Relationship between **A** and **B** is conducted by domain D1. Table **A** and **C** is also connected by domain D1. On the other hand, relationship between **B** and **D** is conducted by D5. Relation among **A**, **B**, **C** and **D** can be also represented by graph as shown in Figure 3.

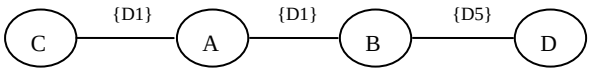


Figure 3. Graph relation of entities

Metadata expressing relation among four tables as given in the example can be simply seen in Table 2.

Table 2. Example of metadata

Table-1	Table-2	Relationship
Table **A**	Table **B**	{D1}
Table **A**	Table **C**	{D1}
Table **B**	Table **D**	{D5}

Through the metadata as given in the example, we may construct six possibilities of denormalized table as shown in Table 3.

Table 3. Possibilities of denormalized tables

No.	Denormalized Table
1	**CA**(D1,D2,D3,D4,D8,D9); **CA**(D1,D2,D8,D9); **CA**(D1,D3,D4,D9), etc.
2	**CAB**(D1,D2,D3,D4,D8,D9,D5,D6,D7), **CAB**(D1,D2,D4,D9,D5,D7), etc.
3	**CABD**(D1,D2,D3,D4,D5,D6,D7,D8,D9, D10,D11,D12), etc.
4	**AB**(D1,D2,D3,D4,D5,D6,D7), etc.
5	**ABD**(D1,D2,D3,D4,D5,D6,D7,D10, D11,D12), etc.
6	**BD**(D5,D6,D7,D10,D11,D12), etc.

CA(D1,D2,D3,D4,D8,D9) means that Table **A** and **C** are joined together, and all their domains are participated as a result of joining process. It is not necessary to take all domains from all joined tables to be included in the result, e.g. **CA**(D1,D2,D8,D9), **CAB**(D1,D2,D4,D9,D5,D7) and so on. In this case, what domains included as a result of the process depends on what domains are needed in the process of mining rules. For D1, D8 and D5 are primary key of Table **A**. **C** and **B**, they are mandatory included in the result, Table **CAB**.

3.2. Table and Function Relation

It is possible for user to define a mathematical function (or table) relation for connecting two or more domains from two different tables in order to perform a relationship between their entities. Generally, the data relationship function performs a mapping process from one or more domains from an entity to one or more domains from its partner entity. Hence, considering the number of domains involved in the process of mapping, it can be verified that there are four possibility relations of mapping.

Let $\mathrm{A}(A_1, A_2, \cdots, A_n)$ and $\mathrm{B}(B_1, B_2, \cdots, B_m)$ be two different entities (tables). Four possibilities of function *f* performing a mapping process are given by:

- One to one relationship
  $f : A_i \rightarrow B_k$
- One to many relationship
  $f : A_i \rightarrow B_{p_1} \times B_{p_2} \times \cdots \times B_{p_k}$
- Many to one relationship
  $f : A_{r_1} \times A_{r_2} \times \cdots \times A_{r_k} \rightarrow B_k$
- Many to many relationship
  $f : A_{r_1} \times A_{r_2} \times \cdots \times A_{r_k} \rightarrow B_{p_1} \times B_{p_2} \times \cdots \times B_{p_k}$

Obviously, there is no any requirement considering type and size of data between domains in **A** and domains in **B**. All connections, types and sizes of data are absolutely dependent on function *f*. Construction of denormalization data is then performed based on the defined function.

4. MULTIDIMENSIONAL ASSOCIATION RULES

As explained in Section 1, association rules that involve two or more dimensions or predicates can be referred to as *multidimensional association rules*. Multidimensional association rules with no repeated predicates are called *interdimension association rules* (e.g. Rule-2)[1]. On the other hand, multidimensional association rules with repeated predicates, which contain multiple occurrences of some predicates, are called *hybrid-dimensional association rules*. The rules may be also considered as combination (hybridization) between intradimension association rules and interdimension association rules. Example of such rule are shown in Rule-3 and Rule-4, the predicate *D* is repeated. Here, we may firstly be interested in mining multidimensional association rules with no repeated predicates or interdimension association rules[7]. Hybrid-dimensional association rules as an extended concept of multidimensional association rules will be discussed later in this paper.

The interdimension association rules may be generated from a relational database or data warehouse with multiple attributes by which each attribute is associated with a predicate. To generate the multidimensional association rules, we introduce an alternative method for mining the rules by searching for the predicate sets. Conceptually, a multidimensional association rule, $A \Rightarrow B$ consists of *A* and *B* as two datasets, called premise and conclusion, respectively.

Formally, *A* is a dataset consisting of several distinct data, where each data value in *A* is taken from a distinct domain attribute in *D* as given by

$$A = \{a_j \mid a_j \in D_j, \text{ for some } j \in \mathrm{N}_n\},$$

where, $D_A \subseteq D$ is a set of domain attributes in which all data values of *A* come from.

Similarly,

$$B = \{b_j \mid b_j \in D_j, \text{ for some } j \in \mathrm{N}_n\},$$

where, $D_B \subseteq D$ is a set of domain attributes in which all data values of *B* come from.

For example, from Rule-2, it can be found that *A*={60, *yes*}, *B*={Lung Cancer}, D_A={*age*, *smoking*} and D_B={*disease*}.

Considering $A \Rightarrow B$ is an interdimension association rule, it can

be proved that $|D_A| = |A|$, $|D_B| = |B|$ and $D_A \cap D_B = \varnothing$. Support of *A* is then defined by:

$$\text{supp}(A) = \frac{|\{t_i \mid d_{ij} = a_j, \forall a_j \in A\}|}{r}, \quad (6)$$

where *r* is the number of records or tuples (see Table 4, *r*=11). Alternatively, *r* in (6) may be changed to $|Q(D_A)|$ by assuming that records or tuples, involved in the process of mining association rules are records in which data values of a certain set of domain attributes, D_A, are not null data. Hence, (6) can be also defined by:

$$\text{supp}(A) = \frac{|\{t_i \mid d_{ij} = a_j, \forall a_j \in A\}|}{|Q(D_A)|} \quad (7)$$

where $Q(D_A)$, simply called *qualified data* of D_A, is defined as a set of record numbers (t_i) in which all data values of domain attributes in D_A are not null data. Formally, $Q(D_A)$ is defined as follows.

$$Q(D_A) = \{t_i \mid d_{ij} \neq null, \forall D_j \in D_A\} \quad (8)$$

Similarly,

$$\text{supp}(B) = \frac{|\{t_i \mid d_{ij} = b_j, \forall b_j \in B\}|}{|Q(D_B)|}. \quad (9)$$

As defined in (3), $\text{support}(A \Rightarrow B)$ is given by

$$\text{supp}(A \Rightarrow B) = \text{supp}(A \cup B) = \frac{|\{t_i \mid d_{ij} = c_j, \forall c_j \in A \cup B\}|}{|Q(D_A \cup D_B)|}, \quad (10)$$

where

$$Q(D_A \cup D_B) = \{t_i \mid d_{ij} \neq null, \forall D_j \in D_A \cup D_B\}$$

$\text{conf}(A \Rightarrow B)$ as a measure of certainty to assess the validity of $A \Rightarrow B$ is calculated by

$$\text{conf}(A \Rightarrow B) = \frac{|\{t_i \mid d_{ij} = c_j, \forall c_j \in A \cup B\}|}{|\{t_i \mid d_{ij} = a_j, \forall a_j \in A\}|} \quad (11)$$

Using the results of (7), (9) and (10), $\text{corr}(A \Rightarrow B)$ is simply calculated by (5).
If supp(*A*) is calculated by (6) and denominator of (10) is changed to *r*, clearly, (10) can be proved having relation as given by (4).

A and *B* in the previous discussion are datasets in which each element of *A* and *B* is an atomic crisp value. To provide a generalized multidimensional association rules, instead of an atomic crisp value, we may consider each element of the datasets to be a dataset of a certain domain attribute. Hence, *A* and *B* are sets of set of data values or sets of datasets. For example, the rule may be represented by

Rule-5:

$$age(X, \text{"20...60"}) \wedge smoking(X, \text{"yes"}) \Rightarrow disease(X, \text{"bronchitis, lung cancer"}),$$

where *A*={{20…29}, {yes}} and B={{bronchitis, lung cancer}}. Simply, let *A* be a generalized dataset. Formally, *A* is given by

$$A = \{A_j \mid A_j \subseteq D_j, \text{ for some } j \in \mathrm{N}_n\}.$$

Corresponding to (7), support of *A* is then defined by:

$$\text{supp}(A) = \frac{|\{t_i \mid d_{ij} \subseteq A_j, \forall A_j \in A\}|}{|Q(D_A)|}. \quad (12)$$

Similar to (10),

$$\text{supp}(A \Rightarrow B) = \text{supp}(A \cup B) = \frac{|\{t_i \mid d_{ij} \subseteq C_j, \forall C_j \in A \cup B\}|}{|Q(D_A \cup D_B)|} \quad (13)$$

Also, $\text{corr}(A \Rightarrow B)$ can be calculated by (5).
Finally, $\text{conf}(A \Rightarrow B)$ is defined by

$$\text{conf}(A \Rightarrow B) = \frac{|\{t_i \mid d_{ij} \subseteq C_j, \forall C_j \in A \cup B\}|}{|\{t_i \mid d_{ij} \subseteq A_j, \forall A_j \in A\}|} \quad (14)$$

To provide a more generalized multidimensional association rules, we may consider *A* and *B* as sets of fuzzy labels. Simply, *A* and *B* are called fuzzy datasets. Rule-3 is an example of such rules, where *A*={*young*, *yes*} and B={*bronchitis*}. Here *young*, *yes* and *bronchitis* are considered as fuzzy lables. A fuzzy dataset is a set of fuzzy lables/ data consisting of several distinct fuzzy labels, where each fuzzy label is represented by a fuzzy set on a certain domain attribute. Let *A* be a fuzzy dataset. Formally, *A* is given by

$$A = \{A_j \mid A_j \in \mathrm{F}(D_j), \text{ for some } j \in \mathrm{N}_n\},$$

where $\mathrm{F}(D_j)$ is a fuzzy power set of D_j, or in other words, A_j is a fuzzy set on D_j.
Corresponding to (7), support of *A* is then defined by:

$$\text{supp}(A) = \frac{\sum_{i=1}^{r} \inf_{A_j \in A} \{\mu_{A_j}(d_{ij})\}}{|Q(D_A)|}. \quad (15)$$

Similar to (10),

$$\text{supp}(A \Rightarrow B) = \text{supp}(A \cup B)$$

$$= \frac{\sum_{i=1}^{r} \inf_{C_j \in A \cup B} \{\mu_{C_j}(d_{ij})\}}{|Q(D_A \cup D_B)|} \quad (16)$$

$\text{conf}(A \Rightarrow B)$ is defined by

$$\text{conf}(A \Rightarrow B) = \frac{\sum_{i=1}^{r} \inf_{C_j \in A \cup B} \{\mu_{C_j}(d_{ij})\}}{\sum_{i=1}^{r} \inf_{A_j \in A} \{\mu_{A_j}(d_{ij})\}} \quad (17)$$

Similarly, if denominators of (15) and (16) are changed to *r* (the number of tuples), (17) can be proved also having relation as given by (4). Here, we may consider and prove that (16) and (17) are generalization of (13) and (14), respectively. On the other hand, (13) and (14) are generalization of (10) and (11).

Finally, $\text{corr}(A \Rightarrow B)$ can be calculated by (5). Alternatively, the correlation betwen two fuzzy datasets can be also defined by the following two definitions.

$$\text{corr}(A \Rightarrow B) = \frac{\sum_{i=1}^{r} \inf_{C_j \in A \cup B} \{\mu_{C_j}(d_{ij})\}}{\sum_{i=1}^{r} \inf_{A_j \in A} \{\mu_{A_j}(d_{ij})\} \times \inf_{B_k \in B} \{\mu_{B_k}(d_{ik})\}} \quad (18)$$

$$\text{corr}(A \Rightarrow B) = \frac{\sum_{t_i \in Q(D_A \cup D_B)} \frac{\inf_{C_j \in A \cup B} \{\mu_{C_j}(d_{ij})\}}{\inf_{A_j \in A} \{\mu_{A_j}(d_{ij})\} \times \inf_{B_k \in B} \{\mu_{B_k}(d_{ik})\}}}{|Q(D_A \cup D_B)|} \quad (19)$$

Here, it depends on the application in which we can use (5), (18) or (19) in order to consider a more matching correlation.

5. HYBRID-MULTIDIMENSIONAL AR

In order to proposed a concept of hybrid-multidimensional association rule, it is necessary to extend the concept of relational data table as described in Section 3 by considering d_{ij} as a dataset ($d_{ij} \subseteq D_j$). Since d_{ij} is a dataset, it is also necessary to define a function $\beta(A,d_{ij})$ representing similarity degree of fuzzy lable *A* given a dataset d_{ij} as follows.

$$\beta(A, d_{ij}) = \frac{\sum_{e \in d_{ij}} \mu_A(e)}{|d_{ij}|} \quad (20)$$

Therefore, (15-19) have to be extended as follows.

$$\text{supp}(A) = \frac{\sum_{i=1}^{r} \inf_{A_j \in A} \{\beta(A_j, d_{ij})\}}{|Q(D_A)|} \quad (21)$$

$$\text{supp}(A \Rightarrow B) = \text{supp}(A \cup B)$$

$$= \frac{\sum_{i=1}^{r} \inf_{C_j \in A \cup B} \{\beta(C_j, d_{ij})\}}{|Q(D_A \cup D_B)|} \quad (22)$$

$$\text{conf}(A \Rightarrow B) = \frac{\sum_{i=1}^{r} \inf_{C_j \in A \cup B} \{\beta(C_j, d_{ij})\}}{\sum_{i=1}^{r} \inf_{A_j \in A} \{\beta(A_j, d_{ij})\}} \quad (23)$$

Similarly, correlation might be calculated by (5) as well as the following equations.

$$\text{corr}(A \Rightarrow B) = \frac{\sum_{i=1}^{r} \inf_{C_j \in A \cup B} \{\beta(C_j, d_{ij})\}}{\sum_{i=1}^{r} \inf_{A_j \in A} \{\beta(A_j, d_{ij})\} \times \inf_{B_k \in B} \{\beta(B_k, d_{ik})\}} \quad (24)$$

$$\text{corr}(A \Rightarrow B) = \frac{\sum_{t_i \in Q(D_A \cup D_B)} \frac{\inf_{C_j \in A \cup B} \{\beta(C_j, d_{ij})\}}{\inf_{A_j \in A} \{\beta(A_j, d_{ij})\} \times \inf_{B_k \in B} \{\beta(B_k, d_{ik})\}}}{|Q(D_A \cup D_B)|}$$

(25)

6. ILLUSTRATIVE EXAMPLE

An illustrative example is given to understand well the concept of the proposed method and how to calculate support, confidence and correlation of the multidimensional fuzzy association rule is performed. The process is started from a given a simple medical records of patients as shown in Table 4.

Table 4. Medical records of patients

Tuples	*Age*	*Smoking*	*Blood Type*	*Diseases*
t_1	20	Yes	A	bronchitis
t_2	25	Yes	A	bronchitis
t_3	22	Yes	B	bronchitis
t_4	27	No	O	diarrhea
t_5	30	No	O	diarrhea
t_6	45	Yes	AB	lung cancer
t_7	40	Yes	O	lung cancer
t_8	50	No	O	diabetes

t_9	60	Yes	B	bronchitis
t_{10}	60	Yes	A	lung cancer
t_{11}	*Null*	No	AB	diarrhea

Based on Table 4, support and confidence of Rule-2 are calculated using (10) and (11), respectively. Related to the conceptual form of the rule $A \Rightarrow B$, it can be followed that *A*={60, yes} and *B*={lung cancer}.

$$\text{supp}(\text{Rule-2}) = \frac{|\{t_{10}\}|}{|\{t_1,...,t_{10}\}|} = 0.1,$$

where $Q(D_A \cup D_B) = \{t_1,...,t_{10}\}$. t_{11} is not included in $Q(D_A \cup D_B)$, because it has a null value in *Ages*. Confidence of Rule-2 is given by

$$\text{conf}(\text{Rule-2}) = \frac{|\{t_{10}\}|}{|\{t_{10}, t_9\}|} = 0.5.$$

Correlation of Rule-2 is given by

$$\text{corr}(\text{Rule-2}) = \frac{\text{supp}(Rule-2)}{\text{supp}(\{60, yes\}) \times \text{supp}(\{bronchitis\})} = \frac{0.1}{0.2 \times 0.4} = 1.25$$

Support, confidence and correlation of Rule-5 are calculated using (13) and (14) as follows.

$$\text{supp}(\text{Rule-5}) = \frac{|\{t_1, t_2, t_3, t_6, t_7, t_9, t_{10}\}|}{|\{t_1,...,t_{10}\}|} = 0.7,$$

$$\text{conf}(\text{Rule-5}) = \frac{|\{t_1, t_2, t_3, t_6, t_7, t_9, t_{10}\}|}{|\{t_1, t_2, t_3, t_6, t_7, t_9, t_{10}\}|} = 1.$$

$$\text{corr}(\text{Rule-5}) = \frac{0.7}{0.7 \times 0.7} = 1.43$$

Rule-3 is a fuzzy rule, where *A*={*young*, *yes*} and B={bronchitis}. *Young* (*yg*) is a fuzzy labels represented by a fuzzy sets as given in Section 1. Support of Rule-3 can be calculated by (16) as shown in the following table.

Table 5. Calculation of fuzzy values

Tuple	$\mu_{yg}(age)$ **α**	$\mu_{ys}(smk)$ **β**	$\mu_{br}(dis)$ **γ**	**min(α,β,γ)**
t_1	1	1	1	1
t_2	0.66	1	1	0.66
t_3	0.87	1	1	0.87
t_4	0.53	0	0	0
t_5	0.33	0	0	0
t_6	0	1	0	0
t_7	0	1	0	0
t_8	0	0	0	0
t_9	0	1	1	0
t_{10}	0	1	0	0
t_{11}	*null*	0	0	0
Σ	3.4	7	4	2.53

Therefore,

$$\text{supp}(\text{Rule-3}) = \frac{2.53}{|\{t_1,...,t_{10}\}|} = 0.253$$

On the other hand, confidence and correlation of Rule-3 are given by

$$\text{conf}(\text{Rule-3}) = \frac{2.53}{2.53} = 1.$$

It can be calculated by (5), (18) and (19) that the correlation is given by

Using (5):

$$\text{corr}(\text{Rule-3}) = \frac{2.53/10}{2.53/10 \times 4/11} = 2.75$$

Using (18):

$$\text{corr}(\text{Rule-3}) = \frac{2.53}{2.53} = 1$$

Using (19):

$$\text{corr}(\text{Rule-3}) = \frac{\frac{1}{1} + \frac{0.66}{0.66} + \frac{0.87}{0.87}}{3} = 1.$$

Positive results of correlations, Rule-1, Rule-5 and Rule-3 show that their conclusion and condition sides are not independent.

6. CONCLUSION

The paper firstly discussed a method of how to provide a denormalized table from a normalized database. Then, a concept of generating multidimensional fuzzy association rules was introduced in the context of mining association rules from medical records of patients. In general, multidimensional association rules consist of two types of rules, namely *interdimension association rules* and *hybrid-dimension association rules*. In this paper, we proposed extended method to generate interdimension association rules as well as hybrid-dimensional association rules. Three sets of equations were introduced to calculate support, confidence and correlation of three different kinds of generalized rules.

7. ACKNOWLEDGEMENT

This work has been supported by the research grant of The Higher Education Directorate of Indonesia (Penelitian Hibah Bersaing) in the years of 2007, 2008 and 2009.

8. REFERENCES

[1] J. Han, M. Kamber, Data Mining: Concepts and Techniques, The Morgan Kaufmann Series, 2001.

[2] G. J. Klir, B. Yuan, Fuzzy Sets and Fuzzy Logic: Theory and Applications, New Jersey: Prentice Hall, 1995.

[3] Rolly Intan, An Algorithm for Generating Single Dimensional Association Rules, Jurnal Informatika Vol. 7 No. 1 (Terakreditasi SK DIKTI No. 56/DIKTI/Kep/2005), May 2006.

[4] Rolly Intan, A Proposal of Fuzzy Multidimensional Association Rules, Jurnal Informatika Vol. 7 No. 2 (Terakreditasi SK DIKTI No. 56/DIKTI/Kep/2005), November 2006.

[5] Rolly Intan, 'A Proposal of an Algorithm for Generating Fuzzy Association Rule Mining in Market Basket Analysis', Proceeding of CIRAS (IEEE). Singapore, 2005

[6] Rolly Intan, 'Generating Multi Dimensional Association Rules Implying Fuzzy Valuse', The International Multi-Conference of Engineers and Computer Scientist, Hong Kong, 2006.

[7] Rolly Intan, Oviliani Yenti, 'Mining Multidimensional Fuzzy Association Rules from a Normalized Database',. Proceedings of International Conference on Convergence and Hybrid Information Technology, IEEE Computer Society, Daejeon, Korea, 2008.

[8] O. P. Gunawan, Perancangan dan Pembuatan Aplikasi Data Mining dengan Konsep Fuzzy c-Covering untuk Membantu Analisis Market Basket pada Swalayan X, (in Indonesian) Final Project, 2004.

[9] L. A. Zadeh, "Fuzzy Sets and systems," International Journal of General Systems, Vol. 17, pp. 129-138, 1990.

[10] R. Agrawal, T. Imielimski, A.N. Swami, "Mining Association Rules between Sets of Items in Large Database", Proccedings of ACM SIGMOD International Conference Management of Data, ACM Press, pp. 207-216, 1993.

[11] R. Agrawal, R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases", Proccedings of 20th International Conference Very Large Databases, Morgan Kaufman, pp. 487-499, 1994.

[12] H. V. Pesiwarissa, Perancangan dan Pembuatan Aplikasi Data Mining dalam Menganalisa Track Records Penyakit Pasien di DR.Haulussy Ambon Menggunakan Fuzzy Association Rule Mining, (in Indonesian) Final Project, 2005.

[13] E.F. Codd, "A Relational Model of Data for Large Shared Data Bank", Communication of the ACM 13(6), pp. 377-387, 1970.

A Context-Based Fuzzy Model for a Generator Bidding System

Moeljono Widjaja
Agency for The Assessment and Application of Technology
BPPT Tower 2, Level 18
Jl. M.H. Thamrin No. 8
Jakarta 10340, Indonesia
moeljono.widjaja@gmail.com

ABSTRACT

This paper proposes a novel approach to strategic bidding in competitive electricity markets that applies a context-based fuzzy model to a generator bidding system. The proposed approach allows the user to modify the weights in the model according to the user's perception of the criteria set in the context. In this approach, Dynamic Programming, which can handle an inter-period optimization of the dispatch of a generator, is applied to the developed fuzzy model to formulate a daily bid of a generator. Practical aspects of formulating the bid are discussed. In addition, the proposed approach is validated by applying it to a market model developed on the market simulator. The performance of the proposed approach is compared to the performance of other techniques such as generic bidding strategies. The validations show that the proposed approach has a significant advantage over generic bidding strategies: it is able to identify a critical market condition where there is an incentive for a generator to exercise its market power.

Keywords

fuzzy fitting, optimization, electricity market, bidding strategy, dynamic programming

1. INTRODUCTION

In a perfectly competitive electricity market, generators have an incentive to offer their electric energy at their respective marginal costs. However, in a market that is not perfectly competitive, generators have some degree of market power that depends on the market conditions, such as the behavior of the competing generators, the load demands and their own cost functions. The exercise of market power in order to maximize profit is also known as strategic bidding.

A strategic bidding can be described as follows: a generator offers its capacity by observing what the market price is and how much it can increase the price by withdrawing some of its capacity from the market such that the price increase outweighs the loss of market share: the profit is then maximized. The strategy depends not only on the price, but also on the rate of price change with respect to the change of the dispatch level.

Theoretically, if the aggregate supply curve of the competitors and the load demand were known exactly, then the optimum solution could be analytically derived [1]. In practice, however, the analytical approach is of very limited use as it over-simplifies the problem [11]. For example, the system usually consists of more than one node. As a result, there is an inter-dependency among market parameters that creates a complex aggregate supply curve which is composed not only of supply curves of local generators but also of incoming and outgoing flows from neighboring nodes. Consequently, the exact aggregate supply curve in a node is analytically difficult to derive.

Game Theory has also been applied in electricity markets as discussed in [18, 4]. Most applications of Game Theory in electricity markets assume non-cooperative games in which a Nash equilibrium exists.

Residual demand analyses have been studied in [7, 5] for optimizing generators' bids. In this approach, a generator optimizes its profit by varying its offer based on its estimated residual demand curve. A residual demand curve is determined by load forecast and expected aggregate supply of competitors. Load forecast can be accurately estimated by a statistical load model developed from its historical data. In contrast, the aggregate supply of competitors is more difficult to estimate since it depends on other market parameters (for example, network constraints and gaming strategies).

Dynamic Programming introduced by Bellman [3] is a useful technique for solving optimization problems involving sequences of decisions such as unit-commitment or generator scheduling. Dynamic Programming was proposed in [6, 12] for formulating bidding strategies in competitive electricity markets.

Applications of heuristic approaches [13, 14, 9] and agent-based model [19] in an electricity market have been reported. Heuristic approaches allow the user's subjective perception to be incorporated in the formulation of bidding strategies. Additionally, they create a user friendly interface between the problem and human beings so that the problem can be

more easily understood by human beings. On the other hand, with the capability of learning from experiences and opponent's behaviors, the agent-based model produces the most viable bidding strategy.

Alternatively, the market behavior can be extracted from historical market data by developing an empirical model. Then, a bidding strategy is formulated based on the developed model. This empirical approach is more robust and flexible than the analytical approach because it can handle a complex system (that is a system with more than one node) and uncertainties. This paper discusses the development of such an empirical approach. Firstly, it presents the development of a context-based fuzzy model for a generator bidding system. Secondly, based on the developed model, it presents the optimization of the generator bidding system using Dynamic Programming. Finally, it validates the proposed approach on a five-node market model.

2. FUZZY MODEL OF A GENERATOR BIDDING SYSTEM

A generator bidding system consists of a controllable input variable such as the dispatch of the generator, non-controllable input variables such as regional load demands and the behavior of competing generators, and a single output variable which is the profit of the generator. Since the actual system is complex, there is a need to develop a representative model that captures the functional relationship between the variables. This section discusses the development of fuzzy model of a generator bidding system from historical market data using the technique discussed in [16].

The development of the model is based on a theoretical analysis of an optimum supply curve. Basically, it states that, for a given load demand, there is an optimum pair of price and dispatch that a generator should offer in order to maximize its profit. Consequently, a context-based fuzzy model is developed to define a fuzzy relationship between dispatch and profit of a generator for a given load demand.

Figure 1 shows the context-based fuzzy model with one input (*Dispatch*), one output (*Profit*), and two context variables (*Load* and *Trading Period*). For each trading period, the input-output relationships are defined by sets of empirical data (the dispatch $\mathbf{s}$ and the profit $\mathbf{\Pi}$) and their corresponding weights as defined by the context. The output membership function is constructed by fitting the weighted output data into a Gaussian-type membership function. The fuzzy relationship between *Dispatch* and *Profit* can be used to maximize the profit of the generator. Once the optimum dispatch is found, the corresponding price can be directly calculated from the profit and the dispatch as given in the following equation:

$$p = \frac{\Pi + C(s)}{s} \quad (1)$$

where p is the price offered by the generator, Π is the profit of the generator, $C(s)$ is the cost function of the generator, and s is the dispatch of the generator.

It turns out that the projection of the input membership function to the output may over-generalize the actual function. Therefore, instead of aggregating the weighted output

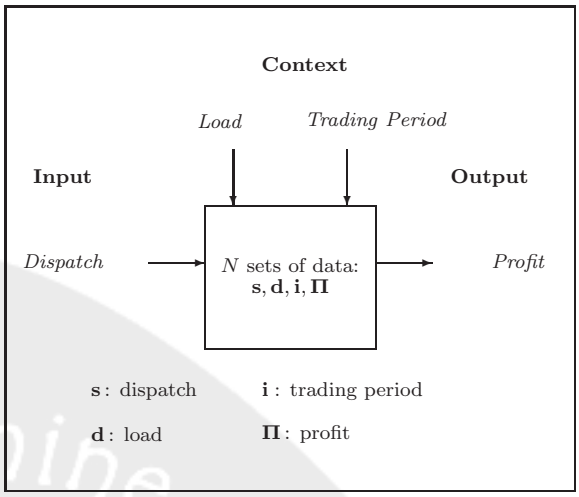


Figure 1: Context-based fuzzy system.

data into a single output membership function, each pair of input-output data with the corresponding weight is treated independently. These weights measure the degree of significance for each data pair. In this model, the input and output variables are in the form of crisp numbers while the context variables are still in the form of fuzzy numbers. This model is used for optimizing a generator bidding system as explained in the next section.

3. OPTIMIZATION OF THE GENERATOR BIDDING SYSTEM

A generator bidding system is optimized by optimizing its fuzzy model that consists of sets of dispatch-profit pairs with their corresponding weights. The structure of the fuzzy model allows the user to modify the weights according to the user's perception of the criteria set in the context. Optimization of bidding strategies involves scheduling of the generator's dispatch over 48 trading periods since the dispatch transition between trading periods is limited by its ramping constraints.

3.1 Bidding Strategies

The fuzzy-based model contains sets of dispatch-profit pairs with corresponding weights defined by the context membership function. These weights indicate the degree of confidence in achieving this profit level at the corresponding dispatch. One of the techniques in modifying these weights is given by the following equation:

$$w_{new} = w^k \quad \text{where} \quad k \geq 0 \quad (2)$$

where w_{new} is the adjusted weight, w is the initial weight set by the context membership function, and k is the exponent. Based on the value of k, there are three possible outcomes:

1. If the exponent k is less than one, then $w_{new} \geq w$. This has the effect of relaxing the criteria set by the context membership function.
2. If the exponent k is equal to one, then $w_{new} = w$. This has no effect on the criteria set by the context

membership function.

3. If the exponent k is greater than one, then $w_{new} \leq w$. This has the effect of strengthening the criteria set by the context membership function.

Based on the subjective perception of the user, the user may choose to relax, keep or strengthen the criteria set by the context membership function. Three types of bidding strategies (Risk-Seeker, Risk-Neutral and Risk-Averse) can be constructed in the fuzzy model by modifying these weights. The three bidding strategies are described as follows:

Risk-Seeker: The user perceives that the criteria set by the context membership function is too stringent; therefore, the criteria is relaxed by setting $k < 1$.

Risk-Neutral: The user perceives that the criteria set by the context membership function is just right; therefore, the criteria is fixed by setting $k = 1$.

Risk-Averse: The user perceives that the criteria set by the context membership function is too relaxed; therefore, the criteria is strengthened by setting $k > 1$.

3.2 Optimization of Generator Dispatch

The dispatch of the generator is optimized by optimizing one of the three bidding models. For each trading period, there are pairs of dispatch and discounted profit. The objective is to maximize the summation of discounted profits over 48 trading periods. It is noted that the dispatch transition between trading periods is limited by the ramping constraints.

The optimization problem is illustrated in the following general example. The objective function is given as:

$$\max \sum_{n=1}^{N} f_n(x_n) \tag{3}$$

subject to:

$$x_{n-1} - x^{Down} \leq x_n \leq x_{n-1} + x^{Up} \tag{4}$$

$$x^{min} \leq x_n \leq x^{max} \tag{5}$$

where x_n is the current state, x_{n-1} is the previous state, $y_n = f_n(x_n)$ is the value corresponding to the state x_n, x^{Up} and x^{Down} are the up/down ramping constraints, x^{min} and x^{max} are the minimum and maximum possible states, and N is the total number of stages.

Two approaches are proposed to optimize the sequences of the state x in order to maximize the summation of y. The first is called a single-period optimization, and the second is an inter-period optimization.

3.2.1 Single-Period Optimization

In a single-period optimization, the "optimum" state is simply chosen to be one of the possible states that gives an immediate maximum return in the current stage. The procedure is illustrated in the following steps:

1. **Initialization of the State:** The initial state is selected to be a state that corresponds to the maximum value in the first stage.
2. **Evaluation of Next Possible States:** The next possible states are evaluated based on the current state and the ramping constraints. They must meet the constraints defined in Equations 4 and 5.
3. **Selection of the Optimum state:** The state with a maximum value, which is one of the feasible states evaluated in the previous step, is selected and Step 2 is repeated.

3.2.2 Inter-Period Optimization

Dynamic Programming introduced by Bellman [3] is a useful technique to solve a class of optimization problems involving sequences of decisions. It has been applied to optimize an economic dispatch of thermal systems [17]. Here, the technique is proposed to optimize dispatch of generators based on the developed bidding model. The implementation of Dynamic Programming is illustrated in the following steps:

1. **Initialization of States:** All unique states in the first stage are selected as possible initial states. If there are two or more identical states, then the state with the maximum value is selected as one of the initial states.
2. **Evaluation of Next Possible States:** The next possible states are evaluated based on the current state and the ramping constraints. They must meet the constraints defined in Equations 4 and 5.
3. **Elimination of Identical States in the Same Stage:** For each sequence (that is a set of n states over n stages), there is a corresponding value which is the summation of value of each state in the sequence. The number of sequences in each stage must be equal to the number of unique states in this stage. However, there is a possibility that more than one sequences reach the same state in the same stage. This means that the number of sequences could be larger than the number of unique states. For each unique state, the sequence with the largest cumulative value is selected, and the remaining sequences are discarded. Basically, this follows the principle of optimality introduced by Bellman [2].
4. **Selection of Optimum Sequence:** In the last stage, the sequence that accumulates the largest value is selected as the optimum sequence.

A direct implementation of Dynamic Programming on all data takes a considerable amount of computing time, especially when the number of data sets is very large. This is also known as "the curse of dimensionality." In order to reduce the dimension of the problem, the data is sampled by partitioning the range of data into several sub-ranges. The determination of the size of the sub-ranges depends on the rates of change up and down. Intuitively, the size of the sub-ranges must be less than or equal to half of the up/down rates of change. This allows feasible paths for the state transition (that is, going up or down) between stages without violating the ramping constraints. In each group, a state with the largest value is selected to represent this group. This guarantees that states in each stage are unique and locally optimum.

4. CASE STUDY

The purpose of this case study is to show the implementation of the proposed approach in a complex market model. The case study was run on an electricity market simulator developed in [15]. A five-node market model was developed imitating the structure in the Australian National Electricity Market. The five regional nodes were QLD, NSW, SNOWY, VIC and SA. Each region except SNOWY had a regional load demand, namely Load QLD, Load NSW, Load VIC and Load SA. For simplification, individual generators in each region were aggregated into five aggregate generators: Generator QLD, Generator NSW, Generator SNOWY, Generator VIC and Generator SA. It is noted that losses in the transmission lines were assumed to be 5% of the power flows. The structure of the model is given in Figure 2. This market model was not meant to be an accurate representation of the Australian NEM. Therefore, the model had some simplifications as explained later in this section.

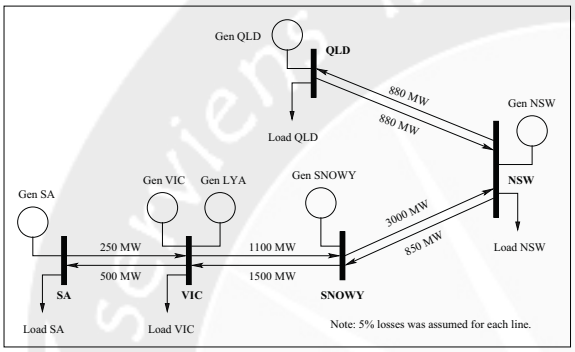


Figure 2: Schematic diagram of the five-node market model (source: [8]).

Load profiles for each region were extracted from ST-PASA which was downloaded from the NEMMCO website. The particular ST-PASA was released on 6 November 2001. It contained seven-day load forecasts for each region in the NEM except SNOWY. The load forecasts for Wednesday (7 November 2001) were used as regional load profiles in this simulation. Figure 3 shows the daily load profiles in Regions QLD, NSW, VIC and SA. It is noted that the magnitude of uncertainty of the load varied over 48 trading periods, especially for the load profiles in Regions QLD and NSW. The market model was developed with the following assumptions. Firstly, it did not model the constraint equations which were used by NEMMCO to operate the power system securely. Secondly, intra-regional loss factors and ancillary services were ignored. In order to compensate for the effects of these assumptions, each regional load was increased by 25%.

Modeling individual generators in the simulator would need a fast computer with large resources. Additionally, running the simulation with a complete and detailed model would take a considerable time as well. In order to reduce the complexity, individual generators in each region were aggregated. The aggregate supply curves in each region were constructed directly from individual supply curves without taking into consideration the intra-regional loss factors. This means that prices offered by individual generators were taken

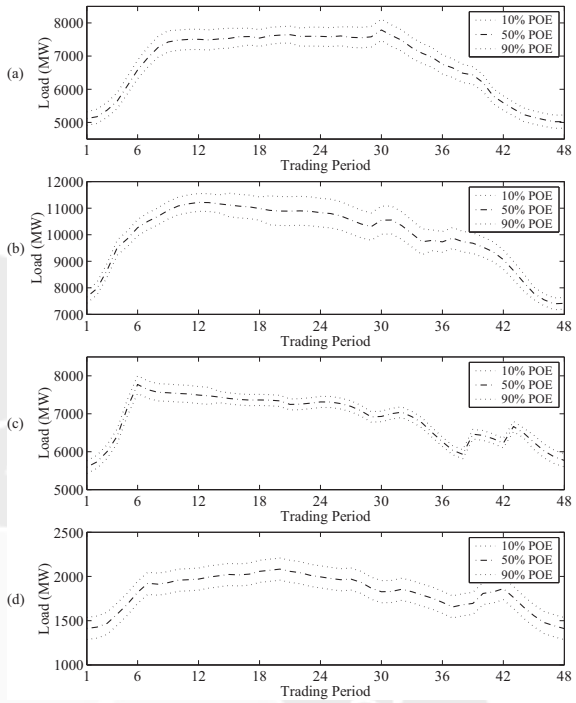


Figure 3: Regional load profiles for the five-node market model: (a) QLD, (b) NSW, (c) VIC and (d) SA.

as they were without converting them to their corresponding prices at regional reference nodes. It is noted that the aggregation ignored the ramp-rate limits on individual generators. This means that the aggregate generators had no ramping constraint.

For Region VIC, the individual generators were aggregated into two aggregate generators. The first was composed of all generators in Region VIC excluding Generators Loy Yang A (unit 1 to 4). The second was an aggregation of the four units of Generator Loy Yang A.

In this example, a generation company (GENCO) Loy Yang A was selected to be the object of the study. This GENCO operated four units of thermal generators in which each unit had a minimum stable load of 320 MW and a maximum capacity of 530 MW. The short run marginal cost for these units was estimated to be around $5.29 per MWh as stated in [10]. The actual cost function is unknown since this is commercially sensitive information. For this model, the four unit thermal generators were aggregated into a single generator, namely Generator LYA. The cost function of this generator was given as:

$$C(s) = 0.006\,s^2 - 10.24\,s + 10000 \text{ for } 1280 \leq s \leq 2120 \quad (6)$$

The ramping constraint for the aggregate generator was 600 MW per 30 minutes, which was simply four times of that of the individual generator (150 MW per 30 minutes).

Three main types of bidding strategies were investigated for this market model.

- The first was a generic bidding strategy with constant bid multipliers: $k = 0.5$, $k = 1.0$, $k = 2.0$ and $k = 3.0$.
- The second was based on a random bidding strategy where Generator LYA offered its electric energy at a random quantity for each trading period.
- The third was a fuzzy-based bidding strategy. The fuzzy-based approach applied three bidding strategies (risk-seeker, risk-neutral and risk-averse) on the model. The resulting pairs of price and quantity formulated using these models are shown in Figure 4. The corresponding daily offers of Generator LYA, which were based on the Cournot supply function.

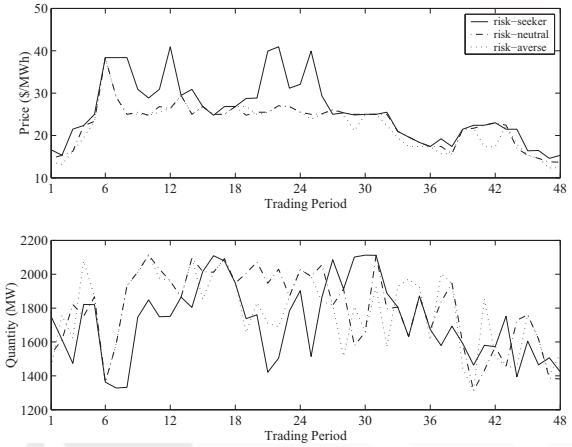


Figure 4: Prices and quantities offered by Generator LYA on a five-node market model based on the following bidding models: risk-seeker, risk-neutral and risk-averse.

Each simulation was run for 100 trading days. The results of the simulation are given in the form of box plots in Figure 5. For generic bidding strategies, the best response was given by a generic bidding strategy with the bid multiplier $k = 2.0$ which gave an average daily profit of $652,247. The random bidding strategy yielded an average daily profit of $631,181. The fuzzy-based approach based on a risk-neutral model gave a better response (that is, an average daily profit of $659,408) than the generic bidding strategy. The spreads of quartile percentages of the daily profits in Figure 5 clearly show the best response of the fuzzy-based approach based on a risk-neutral model compared to other approaches.

5. CONCLUSIONS

This paper has presented the development of a fuzzy-based bidding strategy, where a context-based fuzzy model is developed from empirical data, for a generator participating in a competitive electricity market. The structure of the fuzzy model allows the user to apply different bidding strategies (for example, risk-seeker, risk-neutral and risk-averse) based on the user's subjective perception. The optimization of bidding strategy is carried out using Dynamic Programming on the developed model. The results of the optimization are 48 pairs of optimum dispatch and price. These optimum pairs

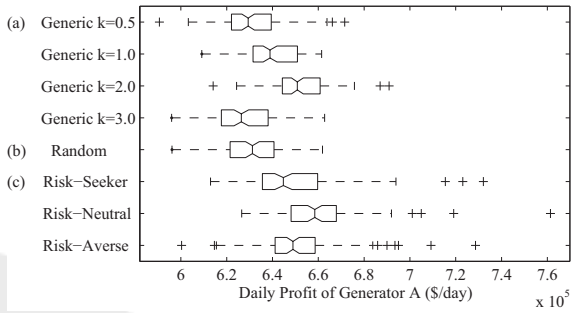


Figure 5: Box plot of daily profits of Generator LYA on a five-node market model using the following bidding strategies: (a) generic bidding, (b) random bidding, and (c) fuzzy-based approaches.

are used to formulate daily optimum supply functions based on the Cournot model.

The proposed bidding strategy, along with generic bidding strategies, was validated on a five-node market models developed on the market simulator. The validations show that the proposed approach has a significant advantage over generic bidding strategies: it is able to identify a critical market condition where there is an incentive for a generator to exercise its market power. The fuzzy-based approach applies Dynamic Programming to the inter-period optimization of dispatch of a generator. The advantage of the proposed fuzzy approach over commonly used methods is the nature of this approach in processing data in the form of fuzzy numbers, which are more tolerant to imprecision. Although the proposed approach is especially tailored to the Australian market, it would also be applicable, with some modifications, to other market structures.

Despite its promising results, there are some minor limitations to the proposed approach. Firstly, there is still a need to refine the model in order to improve the robustness of the developed fuzzy model, especially for complex market models. Secondly, the proposed approach is based on the assumption that the bidding behaviors of the competitors are accurately modeled for each trading period. A more robust approach would need to integrate the bidding extraction technique into the proposed fuzzy model in which the context "bidding pattern" replaces the context "trading period."

6. REFERENCES

[1] E. Anderson and A. Philpott. Optimal offer construction in electricity markets. *Mathematics of Operations Research*, 27(1):82–100, 2002.

[2] R. Bellman. *Dynamic Programming*. Princeton University Press, Princeton, 1957.

[3] R. Bellman and S. E. Dreyfus. *Applied Dynamic Programming*. Princeton University Press, Princeton, 1962.

[4] C. A. Berry, B. F. Hobbs, W. A. Meroney, R. P. O'Neill, and W. R. J. Stewart. Analyzing strategic bidding behavior in transmission networks. *IEEE Tutorial on Game Theory Applications in Electric*

Power Markets, pages 7–32, 1999.

[5] J. Contreras, O. Candiles, J. I. d. l. Fuente, and T. Gomez. A cobweb bidding model for competitive electricity markets. *IEEE Transactions on Power Systems*, 17(1):148–53, February 2002.

[6] R. W. Ferrero and S. M. Shahidehpour. Dynamic economic dispatch in deregulated systems. *Electrical Power & Energy Systems*, 19(7):433–9, 1997.

[7] J. Garcia, J. Roman, J. Barquin, and A. Gonzalez. Strategic bidding in deregulated power systems. In *Proceedings of the 13th Power Systems Computation Conference*, pages 258–264, Trondheim, Norway, 1999.

[8] T. George and P. Ravalli. Operating the national electricity market. *Innovation in Technology for the Electric Power Industry: ESAA 2002 Residential School in Electrical Power Engineering*, 1, 2002.

[9] Y. Y. Hong, S. W. Tsai, and W. M. T. Bidding strategy based on artificial intelligence for a competitive electric market. *IEE Proceedings Generation, Transmission and Distribution*, 148(2):159–64, 2001.

[10] IRPC. Irpc stage 1 report: Proposed sni interconnector. Technical report, Australia, October 2001.

[11] E. P. Kahn. Numerical techniques for analyzing market power in electricity. *The Electricity Journal*, 11(6):34–43, July 1998.

[12] C.-A. Li, A. J. Svoboda, X. Guan, and H. Singh. Revenue adequate bidding strategies in competitive electricity markets. *IEEE Transactions on Power Systems*, 14(2):492 – 497, 1999. read 14 Sept 99.

[13] G. B. Sheble. Decision analysis tools for genco dispatcher. *IEEE Transactions on Power Systems*, 14(2):745–50, May 1999.

[14] H. Song and C.-C. Liu. Future aspects of modern heuristics applications to power systems. In *Proceedings of the 2000 IEEE Power Engineering Society Summer Meeting*, volume 2, pages 1307–11, Singapore, 2000.

[15] M. Widjaja, R. E. Morrison, and L. F. Sugianto. Electricity market simulator using matlab. *Journal of Electrical and Electronics Engineering Australia*, 4(1):77–83, 2002.

[16] M. Widjaja and L. F. Sugianto. Context-based fuzzy system for optimization. In *Proceedings of the 11th IEEE International Conference on Fuzzy Systems*, pages 104–109, Honolulu, Hawaii, USA, 2002.

[17] A. J. Wood and B. F. Wollenberg. *Power generation, operation, and control*. Wiley, New York, 2nd edition, 1996.

[18] Z. Yang, Y. Song, R. Cao, and G. Tang. Analysis on bidding strategy of power provider by game theory. In *Proceedings of 2006 International Conference on Power System Technology*, volume 2, pages 1–6, Chongqing, China, 2006.

[19] Z. Zhang and G. Ma. Strategic bidding model for power generation company based on repast platform. *Journal of Systems Science and Information*, 6(4):381–388, 2008.

Neural Networks for Air-Conditioning Objects Recognition in Industrial Environments

E. Dominguez
University of Malaga
Dept. of Computer Science
Campus Teatinos s/n – 29071 Malaga
enriqued@lcc.uma.es

J. J. Carmona
Altra Çorporacion Empresarial S.L.
PTA - C/. Marie Curie nº 21
29590 - Malaga
jcarmona@altracorporacion.es

ABSTRACT

This paper describes a common recognition problem, related to an industrial environment which has a determined number of different objects that are needed to be recognized. The manufacturing environment is characterized by rapid change, originating new challenges and problems to the production and operation manager in the industry. In response to the need for fast and flexible manufacturing, increasing attention is being given to integration of computing technologies with the manufacturing systems leading to the development of fast and flexible manufacturing systems aided with high performance vision capabilities. In such difficult environments, where objects to be recognized can be dirty and illumination conditions cannot be sufficiently controlled, the required accuracy and rigidity of the system are critical features. Our approach and proposal is based on neural networks. The system works with the bi-dimensional images of the object which are processed briefly before the recognition step. The purpose of the system is to recognize air-conditioning objects for avoiding erroneous identifications due to a large variety of size and kinds of objects. Experimental results on inspection and recognition of a large variety of air-conditioning objects are provided to show the performance of the different network architectures studied.

Keywords

Feedback networks, object recognition, industrial applications.

1. INTRODUCTION

Object recognition is the goal of many computer vision and image analysis applications. Many ways have been explored and proposed such as textured-based systems or color-based systems, being shape-based the most common and dominant. Object recognition in noisy and cluttered scenes is a challenging problem in computer vision. There has been extensive research in the area of computer vision, both in the academic and industrial sector. The drive for this trend is towards greater efficiency and flexibility in production. Effective and successful algorithms and systems have already been reported, some of which are recorded in the manufacturing sector. The manufacturing environments are characterized by rapid change, posing new problems to the production and operations manager in the industry. Using human inspectors for these tasks, it is almost impossible to achieve 100% product quality control for high rates of production. Therefore, machine vision systems may be regarded as a complement to human operators because of their efficiency and accuracy. Under these circumstances, process flexibility is becoming a major priority for many organizations as they attempt to deal with these changes. In response to the need for process flexibility, increasing attention is being given to integration of computing technologies with the manufacturing systems leading to the development of flexible manufacturing systems aided with high performance vision capabilities.

The proposed object recognition system is based on neural networks and the shape of the air-conditioning objects. Usually, in a typical machine vision shape-based system a characteristics vector was calculated from the image, once the shape has been extracted from the background. In industrial environments where the single production is manufactured, there is clear need to fast recognition of objects. In this sense, the proposed system works with images instead with characteristics vectors to make the process computational lighter. Therefore, when the shape of the object is extracted from the background no further calculations are needed, and the image is ready for the recognition process.

The aim of the paper is to present an outline of a neural network for air-conditioning objects recognition (fig. 1). The proposed neural model consists of a feedforward network that identifies (the shape of) air-conditioning objects in an industrial environment. The paper discusses mainly several training algorithms that are implemented and tested.

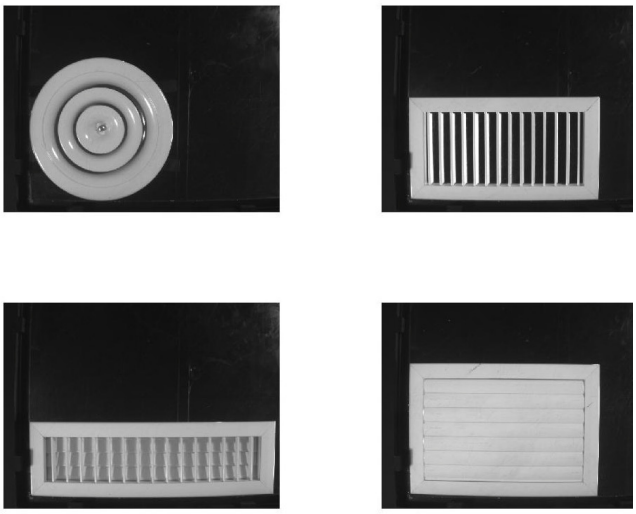


Figure 1. Examples of air-conditioning objects.

In the next section a brief description of the image processing is provided. Section 3 presents the proposed network architecture and the training algorithms implemented. The performances of the

different neural models are shown in section 4. Finally, concluding remarks are drawn in section 5.

2. IMAGE PROCESSING

The proposed neural recognition system is included into a whole industrial system, which was implemented for object recognition and inspection. Although the whole system is composed by five cameras, this paper is only focused on the object recognition from the bottom camera which is the unique camera used for recognition. The rest of cameras are used to inspect different parts of the air-conditioning objects.

In a difficult industrial environment the object recognition is a hard task due to the noisy and cluttered scenes, moreover when the recognition must do it in real time. Therefore, the image preprocessing is an important task to obtain good results in the recognition. Most of the industrial recognition process are unable to identify anything without a preprocessing task. This is a drawback in the real-time systems, since the time needed to preprocess the image is subtracted from the response time of the system. Consequently, the available time for the recognition task is reduced. An advantage of the proposed neural recognition system is that the preprocessing task is eliminated, then the response time of the system is spent in the recognition task.

Before the image is processed by the neural network, a previous phase is necessary to avoid the typical noise in industrial environments. In order to identify the kind of the air-conditioning objects, a subimage capturing the shape of the object is presented as input of the neural network. Therefore, the final objective of the image processing is to obtain a 20x30 pixel binary window centered in the object.

The image processing is performed in two steps: The first step locates the object and captures the shape. In the next step the window centered in the object is determined.

2.1 Preprocessing

Due to the dirty of industrial environment a noise removal filter and image enhancement transformations must be applied before segmentation. For dealing with noise, a gaussian filter is applied before the location of the object and segmentation.

2.2 Segmentation

The segmentation of the object is based on the image histogram and the shape using the boundaries of the silhouette image. Silhouettes are limited fundamentally as shape descriptors for general objects, since they ignore internal contours essentials for the identification. Early works using Fourier descriptors [4][5] or transformation capturing the structure of the shape [6] are found in the literature. However, the consuming time of these algorithms is prohibited in the real-time context.

A smooth function (1) is applied before the threshold is calculated from the image histogram. This function replaces the original value of the histogram in the point b by the average of a window given by W, centered in b.

$$h_{smooth}[b] = \frac{1}{W}\sum_{w=-(W-1)/2}^{(W-1)/2} h_{raw}[b-w] \quad (1)$$

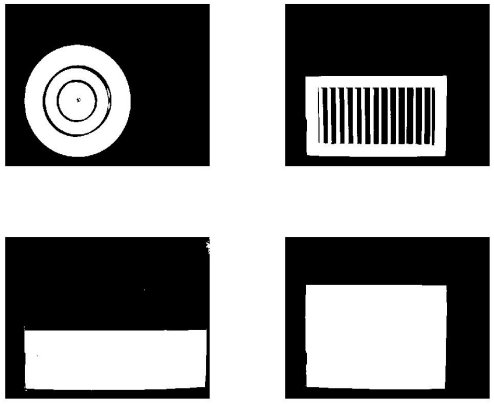


Figure 2. Object segmentation

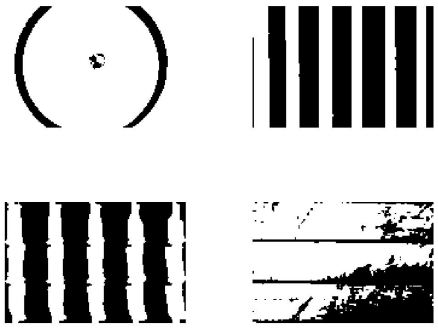


Figure 3. 20x30 binary images (ROIs)

Once the histogram has been smoothered, the global maximum in the histogram G_m is selected, and then the grey level G where $h(G)|G_m-G|$ is maximum. The threshold will be the minimum of the histogram between G and G_m. This is not an optimum threshold calculation, but in this step the goal is to extract the object (fig. 2), without worrying for loosing objects details.

2.3 Selecting the ROI

After the segmentation process, the image contrast is adjusted in order to cover the full range between the zero and one grey values.

A new threshold is calculated by the Otsu method. The region of interest (ROI) is a 20x30 binary image (fig. 3), which eliminates redundant information without loosing the object structure information. This reduction of information makes the network much more light and fast to train.

2.4 Adaptability

Due to the traditional changes in the industrial environments, the adaptability is an essential characteristic in order to remain the effectiveness and suitability of the proposed system. Problems related to dust, noise and small vibrations and movements during the images capture are solved raising the robustness of the system.

Consequently, an adaptive process is established consisting of a new training phase is activated by taking thousands of stored images from known air-conditioning objects and performing a supervised analysis. This process is eventually done when the system is not in use or idle. Moreover, this adaptive process is also performed due to the manufacture of new objects by the company. The images of the new manufactured objects are previously

included into the training set of the system as a supervised task. Note that this adaptive process causes an increment of the system flexibility and scalability.

3. NEURAL MODEL

Feedback neural network architecture is considered in this work. Multiple training algorithms are tested, due to their differences in accuracy results and computational performance.

The neural architecture consists of two hidden layers, the first one with 600 inputs (one per pixel of the ROI), and an output layer formed by *m* neurons, where *m* is the number of kinds of air-conditioning objects to recognize. In general, let N_k be the number of neurons at the *k*-th layer and s_i^k the output of the *i*-th neuron at the *k*-th layer, the computational dynamics is defined as follows

$$s_i^k = f\left(\sum_{j=1}^{N_{k-1}} w_{ij}^k s_j^{k-1}\right) \quad (2)$$

where *f* is the transfer function and *w* is the synaptic weight. The three layers have a log-sigmoid function as transfer function, which allows only output values in each layer between zero and one.

3.1 Learning

Several training algorithms are compared in order to analyze the training process and the size of the hidden layers. An extensive description of different training algorithm can be found in [1-3]. Below a brief description is provided with the essentials characteristics of the tested training algorithms.

3.1.1 Variable learning rate backpropagation (VLR)

The learning rate is increased in the current epoch of the training if the new error exceeds the old error (more than a defined parameter) or decreased if the new error is less than the new error.

3.1.2 Variable learning rate backpropagation with momentums (VLRM)

This algorithm combines the last one with momentum training. Momentum allows the network to ignore small features on the error surface

3.1.3 Resilient backpropagation (RPROP)

The sigmoid functions used in the network have a slope that approaches to zero when the input gets large. This can be a problem when using steepest descent to train the network since the magnitude of the gradient can be very small and so will be changes in weights and biases. In the resilient backpropagation algorithm [7] the magnitude of the derivative has no effect in the weight update.

3.1.4 Powell Beale restarts (PBR)

As in all conjugate gradient algorithms, the search direction is periodically reset to the negative of the gradient. In this algorithm [8], the search direction is reset when there is very little orthogonality left between the current gradient and the previous gradient.

3.1.5 Scaled conjugate gradient (SCG)

Standard conjugate gradient algorithm [9] must perform a line search at each iteration, and this search is computationally expensive. To avoid this line search, this algorithm combines the model-trust region approach with the conjugate gradient approach.

4. EXPERIMENTAL RESULTS

To evaluate the performance of the different training algorithms, a set of 152 air-conditioning object images is prepared for the training of the network. Another 34 images are prepared in an additional validation set. This validation set will allow us to control the generalization of the network.

A conventional computer (Pentium IV 1.6 GHz with 1Gb RAM) has been used for the experiments. The neural networks were simulated on this platform using Matlab code.

Table 1. Variable learning rate backpropagation matching results

Layer size		Training epochs	Matching results	
First	Second		Training set	Validation set
170	100	264	99.34%	88.23%
170	120	287	98.03%	91.18%
150	120	300	100%	91.18%
170	130	216	98.03%	88.23%
170	150	296	99.34%	91.18%
190	130	228	100%	85.29%
190	150	300	100%	88.23%

Table 2. Variable learning rate backpropagation with momentums matching results

Layer size		Training epochs	Matching results	
First	Second		Training set	Validation set
170	100	227	99.34%	88.23%
170	120	199	97.37%	94.12%
150	120	300	99.34%	85.29%
170	130	208	98.03%	85.29%
170	150	222	100%	91.18%
190	130	198	98.02%	88.23%
190	150	191	96.05%	88.23%

Table 3. Resilient backpropagation matching results

Layer size		Training epochs	Matching results	
First	Second		Training set	Validation set

170	100	38	100%	91.18%
170	120	36	99.34%	91.18%
150	120	36	99.34%	94.12%
170	130	37	100%	91.18%
170	150	36	99.34%	88.23%
190	130	42	99.34%	91.18%
190	150	39	100%	94.12%

Table 4. Powell Beale restarts matching results

Layer size		Training epochs	Matching results	
First	Second		Training set	Validation set
170	100	68	98.03%	91.18%
170	120	74	99.34%	91.18%
150	120	53	99.34%	88.23%
170	130	58	99.34%	82.53%
170	150	48	99.34%	88.23%
190	130	61	94.08%	79.41%
190	150	61	98.68%	91.18%

Table 5. Scaled conjugate gradient matching results

Layer size		Training epochs	Matching results	
First	Second		Training set	Validation set
170	100	97	100%	91.18%
170	120	76	99.34%	82.35%
150	120	86	98.03%	88.23%
170	130	69	98.03%	88.23%
170	150	61	100%	88.23%
190	130	90	99.34%	91.18%
190	150	69	99.34%	85.29%

Table 6. Variable learning rate backpropagation time results

Layer size		Training time (sec)	Training set (sec)	Validation set (sec)
First	Second			
170	100	30.26	0.782	0.843
170	120	38.17	0.777	0.841
150	120	40.43	0.782	0.841
170	130	33.70	0.780	0.841
170	150	37.5	0.775	0.833
190	130	40.40	0.779	0.838
190	150	49.60	0.777	0.838

Table 7. Variable learning rate backpropagation with momentums time results

Layer size		Training time (sec)	Training set (sec)	Validation set (sec)
First	Second			
170	100	31.09	0.777	0.837
170	120	30	0.779	0.835
150	120	42.34	0.777	0.835
170	130	31.76	0.775	0.835
170	150	32.76	0.775	0.834
190	130	33.31	0.778	0.838
190	150	34.45	0.778	0.836

Table 8. Resilient backpropagation time results

Layer size		Training time (sec)	Training set (sec)	Validation set (sec)
First	Second			
170	100	6.891	0.778	0.837
170	120	6.844	0.777	0.839
150	120	6.172	0.779	0.837
170	130	7.25	0.778	0.836
170	150	7.281	0.782	0.840
190	130	9.062	0.781	0.840
190	150	8.891	0.781	0.843

Table 9. Powell Beale restarts time results

Layer size		Training time (sec)	Training set (sec)	Validation set (sec)
First	Second			

170	100	24.22	0.779	0.840
170	120	30.53	0.781	0.846
150	120	18.07	0.781	0.842
170	130	24.17	0.782	0.839
170	150	19.75	0.780	0.842
190	130	28.07	0.777	0.842
190	150	33.68	0.782	0.844

Table 10. Scaled conjugate gradient time results

Layer size		Training time (sec)	Training set (sec)	Validation set (sec)
First	Second			
170	100	27.12	0.780	0.840
170	120	21.75	0.782	0.843
150	120	22.97	0.778	0.843
170	130	21.15	0.776	0.841
170	150	18.76	0.780	0.842
190	130	28.82	0.781	0.842
190	150	24.34	0.778	0.841

According the tests carried out in various illumination conditions and camera settings in the actual system, the different air-conditioning objects was recognized successfully in many cases. In the experiment results, however, there were some unsuccessful cases also.

Performance results in classification accuracy and information about the training process are shown in the tables 1-5. The resilient backpropagation algorithm (RPROP) shows promising results and suitability for the chosen neural architecture.

Experiment results show that small modifications of the size of hidden layers are not significant in the performance of the tested training algorithms. These modifications are more significant in the two variable learning rate algorithms (VLR and VLRM). Although the percentage of good recognition is similar in all tested training algorithms, the number of training epochs is very different between them.

A time comparison is provided by the tables 6-10. The resilient backpropagation algorithm shows clearly more suitability in a real-time environment due to its lower computation time. Although the results are far from being definitive, since a 100% is not guaranteed, an in-depth study of the best training algorithms (adjusting them to this specific problem) is a promising further work. The goal is to achieve a performance as nearest to 100% as possible adding an unknown kind of objects to avoid erroneous identification.

5. CONCLUSIONS AND FUTURE WORKS

In this paper a neural system for air-conditioning objects recognition in difficult industrial environments is provided. This system can also be applied for recognizing other kinds of objects, since a generic feature and brightness based method is used. One of the main advantages of this kind of methods is the fast and flexible recognition of objects. In manufacturing environments, the requirements for flexibility, speed and reliability are high. Therefore, most theoretical methods with high accuracy cannot be applied due to the time-consuming computation.

In our study, different training algorithms have been tested in order to compare and analyze the importance of the learning process in a neural system. Moreover, different neural architectures have been considered modifying the size of the hidden layers.

A key characteristic of the neural system is the estimation of shape similarity based on a simple image processing. Due to the neural network is able to learn the shape of the different objects; the image processing is minimum and practically unnecessary. Therefore, the neural approach is simple and much easier to apply and implement than the general and complicated statistical methods.

According the experiment results, the proposed neural system can be used for recognizing air-conditioning objects with a tiny human supervision, since the 100% is not always guarantied. In future work, we intend to add an unknowing kind of object in order to avoid erroneous identification.

6. REFERENCES

[1] Patterson D.W., “Artificial Neural Networks: Theory and Applications”, Prentice Hall

[2] L.P.J. Veelenturf, “Analysis and Applications of Artificial Neural Networks”, Prentice Hall 1995

[3] Christopher M. Bishop, “Neural networks for pattern recognition”, Clarendon Press 2000

[4] C. Zahn and R. Roskies, “Fourier descriptors for plane closed curves,” IEEE Trans. Computers, vol. 21, no. 3, pp. 269-281, 1972.

[5] E. Persoon and K. Fu, “Shape discrimination using Fourier descriptors,” IEEE Trans. Systems, Man. And Cybernetics, vol. 7, no. 3, pp. 170-179, 1977.

[6] D. Sharvit, J. Chan, H. Tek and B. Kimia, “symmetry-based indexing of image databases,” J. Visual Comm. And Image Representation, vol. 9, no. 4, pp. 366-380, 1998.

[7] Riedmiller, M., and H. Braun, “A direct adaptive method for faster backpropagation learning: The RPROP algorithm,” Proceedings of the IEEE International Conference on Neural Networks, 1993.

[8] Powell, M. J. D., “Restart procedures for the conjugate gradient method,” Mathematical Programming, vol. 12, pp. 241-254, 1977.

[9] Moller, M. F., “A scaled conjugate gradient algorithm for fast supervised learning,” Neural Networks, vol. 6, pp. 525-533, 1993.

Pattern Recognition Using Discrete Wavelet Transformation and Fuzzy Adaptive Resonance Theory

Arnold Aribowo
Computer Engineering, Computer Science Faculty, Universitas Pelita Harapan
UPH Tower, Lippo Karawaci, Tangerang, 15811
arnold.aribowo@staff.uph.edu

Samuel Lukas
Informatics Engineering, Computer Science Faculty, Universitas Pelita Harapan
UPH Tower, Lippo Karawaci, Tangerang, 15811
samuel.lukas@staff.uph.edu

Joannes Franciscus
Informatics Engineering, Computer Science Faculty, Universitas Pelita Harapan
UPH Tower, Lippo Karawaci, Tangerang, 15811
j0e2_ne0@hotmail.com

ABSTRACT
With many types and variations of existing patterns and also needs of the high recognition accuracy, a pattern recognition process becomes very important and crucial. Many recognition techniques are developed to reach a similar ability with human capabilities in processing data.

This paper elaborates the design of software to perform pattern recognition by using Wavelet Discrete Transformation and Fuzzy Adaptive Resonance Theory. At the beginning, the system captures the image pattern. Then, the system uses image processing techniques including Wavelet Discrete Transformation to enhance the quality of the image pattern. Finally, the result of transformation is used as an input of the Fuzzy Adaptive Resonance Theory for classifying the pattern.

From the experiments, with the learning parameter value of 1 and vigilance parameter is set to 0.87, the system reached the highest accuracy of the untrained pattern recognition with 91.66% success rate. The system also obtains a perfect match for the trained data using the learning rate of 1 and various vigilance parameters.

Keywords
Discrete Wavelet Transformation, Fuzzy Adaptive Resonance Theory

1. INTRODUCTION
Many research efforts are devoted to recognize variations of patterns to yield high recognition accuracy. Many recognition techniques are developed to reach a similar ability with human capabilities in processing data.

Relying on recognition techniques without considering image processing technique usually results in low success rate. Therefore in this paper, image processing techniques are described. Finally, Fuzzy Adaptive Resonance Theory is used to recognize patterns. In [1,2], fuzzy ART is used for handwritten recognition. After introducing the basic issue of this research, this paper provides the related theoretical backgrounds in section 2. In section 3, the design of the system is provided. Then, the paper gives the result of experiments in section 4. Finally, section 5 presents conclusions of the paper.

2. THEORETICAL BACKGROUNDS
The brief theoretical backgrounds discussed in this paper consist of image processing techniques and Fuzzy ART. A number of techniques are used, including thresholding, rotation, scalling, and Discrete Wavelet Transformation. As translation and rotation needs Freeman Chain Code for determining centre of gravity and contour tracing, Freeman Chain Code technique is also discussed in this section.

2.1 Image Processing
Image Processing is a process to improve an image quality in such away that is easy to interpret by human or computer. An input for this process is image. An output is also image with a better quality.

2.2 Freeman Chain Code
Chain codes are one of the shape representations which are used to represent a boundary by a connected sequence of straight line segments of specified length and direction. This method uses a sequence of code to represent an image. One of this representation is on 8-connectivity of the segments. The direction of each segment is coded by using a numbering scheme as shown in Figure 1 below [4]. Chain codes based from this scheme are known as Freeman chain codes.

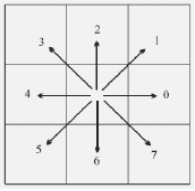


Figure 1. Freeman chain code

Beside its usage to represent an object, freeman chain code is also used to obtain information of an object, such as area, and perimeter. By obtaining freeman chain code of an image, a centre of gravity of that object can also be determined.

2.3 Wavelet Discrete Transformation
A wavelet transform is the representation of a function by wavelets. The wavelets are scaled and translated copies (known as "daughter wavelets") of a finite-length or fast-decaying

oscillating waveform (known as the "mother wavelet"). Wavelet transforms are classified into discrete wavelet transforms (DWT) and continuous wavelet transforms (CWT). The word wavelet is due to Morlet and Grossmann in the early 1980s [5]. Discrete wavelet transformation is able to decompose signal to obtain low-frequency and high-frequency part of a signal. Low frequencies correspond to global information of a signal, whereas high frequencies correspond to detail information of a signal. Three levels of wavelet decomposition can be depicted in the following figure:

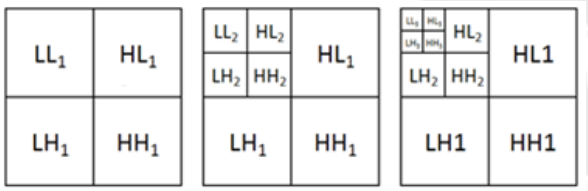


Figure 2. One, two and three levels of wavelet decomposition

2.4 Fuzzy ART

Architecture of fuzzy ART can be shown in the following figure:

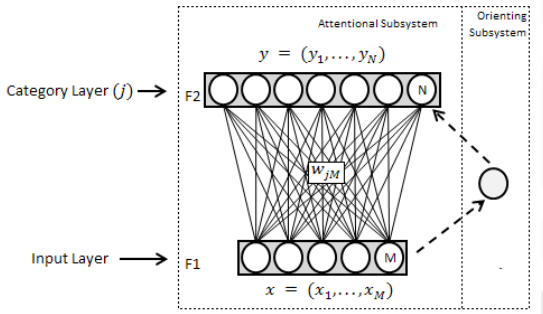


Figure 3. Architecture of Fuzzy ART

According to [3], in fuzzy ART, each input I is an M-dimensional vector(I_1,...,I_M), where each component I_i is in the interval [0, 1]. That input is given to the input layer (F_1) and represented as X_1,…,X_M. Category layer (F_2) with N category is represented as Y_1,…,Y_N. Each category j (j=1,…, N) corresponds to a vector $w_j \equiv (w_{j1},\ldots,w_{jM})$ of adaptive weights, which has initial value of 1. A Fuzzy ART is constructed according to several parameters. They are choice parameter $\alpha > 0$, learning rate parameter $\beta \in [0, 1]$, and vigilance parameter $\rho \in [0, 1]$.

Initially, fuzzy ART get input I which is then stated as X_1,…,X_M. Based on the input, choice function Tj for each category in F_2 is computed. At the beginning, there is one category in F_2. Choice function is defined in the following way:

$$T_j(I) = \frac{|I \wedge w_j|}{\alpha + |w_j|}$$

Where Fuzzy AND $(\wedge)$ operation and $|\cdot|$ is defined as follows:

$$(p \wedge q) = \min(p, q)$$

$$|p| = \sum_{i=1}^{M} p_i$$

Each value of choice function in F_2 enable a competition occurs. Only one choice function in category layer which will be the winner. In the next step, Fuzzy ART gives a value to category layer in the following way:

$$yi = \begin{cases} 0, & \text{for inactive neuron} \\ 1, & \text{for active neuron} \end{cases}$$

Where y_i is an activation vector in F_2 and i is index in F_2.

Fuzzy ART then checks whether the match function of the chosen category meets the vigilance criterion through the following equation :

$$\frac{|I \wedge w_j|}{|I|} \geq \rho$$

If the above equation does not hold, mismatch reset is performed to obtain another category as a winner. This search is continued until there is category as winner which fulfills a resonance condition. If there is no such category, fuzzy ART forms new category and selects it as a winner. When resonance occurs, weight vector is modified according to the following equation:

$$w_{j(new)} = \beta(I \wedge w_{j(old)}) + (1-\beta) w_{j(old)}$$

3. DESIGN OF THE SYSTEM

The system consists of two main parts, training of images data and recognizing them. In the first phase, the image is processed using several image processing methods including discrete wavelet transformation. It transforms data of 64 x 64 pixel and yields image data of 8 x 8 pixel. The result is then normalized in such a way that it can be composed as a training data for Fuzzy ART.

In the testing phase, an input image is processed and then it is classified according to the existing category and weight which is obtained through the training process. Architecture of the system is depicted in the following figure:

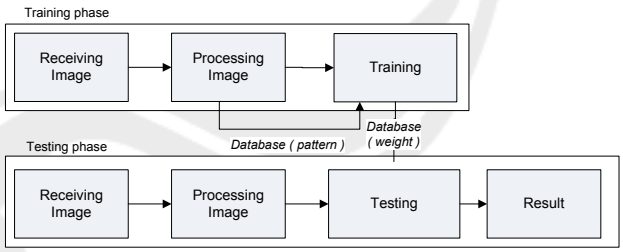


Figure 4. Architecture of the system

The process of receiving and processing an image consists of receiving an input image, followed by several consecutive processes, including thresholding, translation, rotation, cropping, scalling, and discrete wavelet transformation.

The process of fuzzy ART training can be shown in the following way:

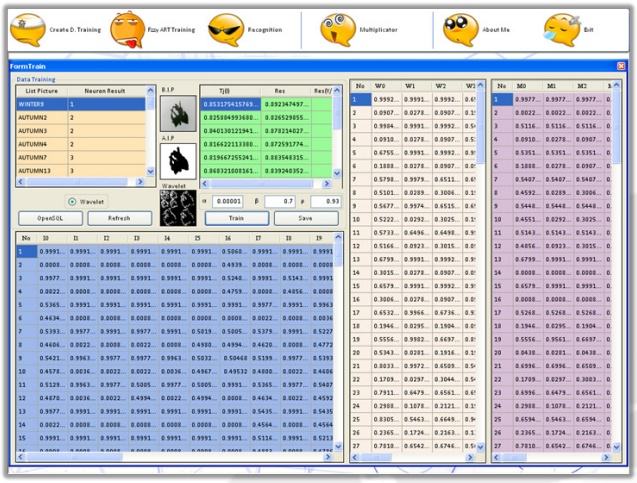


Figure 5. Training of Fuzzy ART

The process of fuzzy ART testing can be depicted as follows:

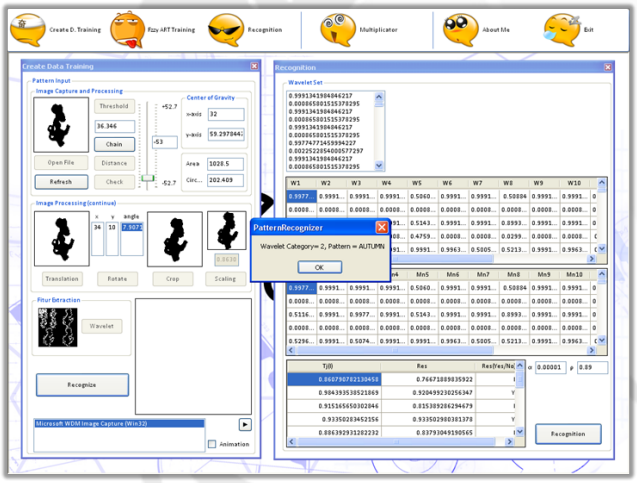


Figure 6. Testing of Fuzzy ART

4. EXPERIMENTAL RESULT

In the testing phase, 6 images are mainly used. For each image, variations of 15 patterns are applied, and therefore 90 images are utilized as a sample data. During the training process, 5 patterns of each image are used, and therefore 30 images are used for training process. In this experiment, the fixed choice parameter (α) of value …is applied. The varying learning rate (β) of value 0.6, 0.7, 0.8, 0.9 and 1 are used. The varying vigilance parameter (ρ) of value between 0.87 and 0.92 and of value 1 are set.

4.1. Experiment On The Amount Of Created Categories

The relation between the amount of created categories and varying vigilance parameter (ρ) and learning rate (β) is depicted in the following figure:

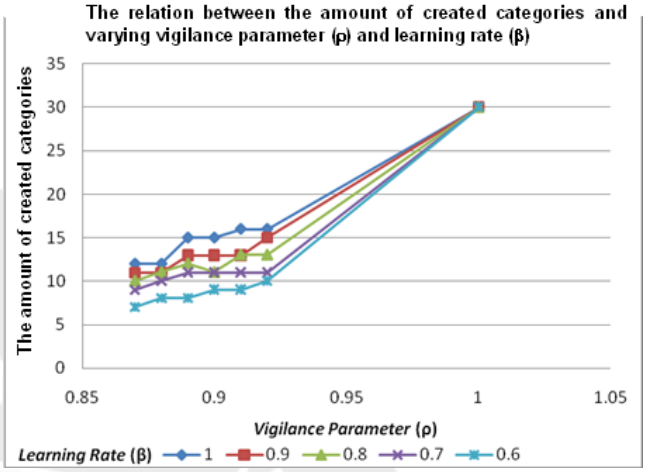


Figure 7. The relation between the amount of created categories and varying vigilance parameter and learning rate

From the above figure, it can be shown that the higher the value of vigilance parameter, the more possibility of mismatch reset and therefore new categories are added.

4.2. Accuracy Rate Of Trained Data Samples Recognition

In this testing, 30 data samples are used. The accuracy of trained data samples for varying vigilance parameter and learning rate is presented in the following graph:

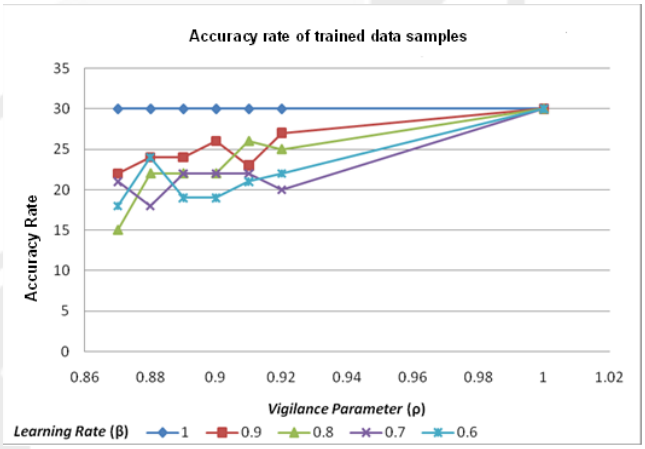


Figure 8. Accuracy rate of trained data samples

From the above figure, it can be shown that for 30 data that have been trained, accuracy rate of recognitions are 100% with learning rate (β) is of value 1 and various vigilance parameter (ρ).

4.3. Accuracy Rate Of New Data Samples Recognition

In this testing, 60 data samples are used. The accuracy of new data samples for varying vigilance parameter and learning rate is presented in the following graph:

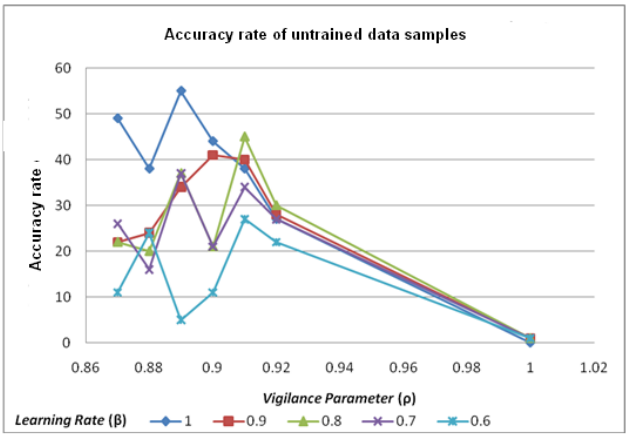


Figure 9. Accuracy rate of untrained data samples

From the above figure, it can be shown that the highest accuracy rate of recognitions are 91.66 % with learning rate (β) is of value 1 and vigilance parameter (ρ) is of value 0.89.

5. CONCLUSION

To summarize the software for recognizing patterns was developed by implementing discrete wavelet transformation and fuzzy ART. According to several testing for 30 data that have been trained, it can be shown that accuracy rate of recognitions are 100% with learning rate (β) is of value 1 and various vigilance parameter (ρ). From several testing for 60 new data samples, it can be shown that the highest accuracy rate of recognitions are 91.66 % with learning rate (β) is of value 1 and vigilance parameter (ρ) is of value 0.89. In the future, more advanced testing can be performed with various choice parameters (α). More image processing techniques can also be added to improve quality of images.

6. REFERENCES

[1] Arnold Aribowo, Samuel Lukas, Andre Tirta Winoto, 2009, The Design and Implementation of Data Entry Automation Software Using Fuzzy ARTMAP, Proceedings International Industrial Informatics Seminar 2009, UIN Yogyakarta, 15 Agustus 2009.

[2] Chergui Leila, Benmohammed Mohammed, 2008, ART Network for Arabic Handwritten recognition system, http://eref.uqu.edu.sa/files/eref2/folder6/f78.pdf.

[3] Gail A. Carpenter, S. Grossberg, dan David B. Rosen., 1991, Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system. Oxford UK: Elsevier Science Ltd.

[4] Gonzalez, R.C., Woods, R.E., 1992, Digital Image Processing 3rd edition, Addison-Wesley.

[5] P. Robi, , 1996, FUNDAMENTAL CONCEPTS & AN OVERVIEW OF THE WAVELET THEORY, Second Edition. Dept. of Electrical and Computer Engineering. Glassboro: Rowan University.

Resolving Occlusion in Multi-Object Tracking Using Fuzzy Similarity Measure

Rahmatri Mardiko
Faculty of Computer Science University of Indonesia
Kampus UI Depok 16424 - Indonesia
+62 21 786 3419
mardiko@gmail.com

M. Rahmat Widyanto
Faculty of Computer Science University of Indonesia
Kampus UI Depok 16424 - Indonesia
+62 21 786 3419
widyanto@cs.ui.ac.id

ABSTRACT
Occlusion is a very common problem in multi-object tracking for real world scenes. In this paper, a method for occlusion handling based on fuzzy similarity measure is proposed. Fuzzy approach has been successfully used in many fields where there exists uncertainty and imprecision. In the occlusion problem, similarity measure is incorporated to resolve track of an object after occlusion happens. Here, the fuzzy similarity measure is done by representing color, texture, and shape of objects as fuzzy features and using Fuzzy Feature Contrast Model (FFCM) with its derivatives to perform similarity measure. Experimental results show that the proposed method can handle occlusion effectively with moderately fast computational time.

Keywords
Multi-object tracking, occlusion handling, fuzzy similarity measure.

1. INTRODUCTION
Automated visual surveillance is an emerging field within computer vision community which has attracted many attentions for researcher. Its aim is to obtain high level description about the captured scenes in the video. To achieve this goal, several processes are required including environment modeling, foreground extraction, object classification, tracking, and then high level analysis [1]. Environment modeling is a process to acquire static components of the scene (background) while foreground extraction attempts to get the dynamic parts (foreground). The extracted foreground is then recognized as object. Classification is performed to distinguish between different types of objects such as human, car, animals, etc. During object appearance in the video, its trajectory is recorded by tracking process which is done by object correspondence in consecutive frames. Finally, ones can get the description by analyzing object movement, trajectory, or recognize its identity, behavior, and so on.

In real world scenes, the objects are frequently cluttered and interactions among them are inevitable. It can degenerate tracking accuracy as it is difficult to recognize individual object correctly when they are occluded each other. This problem is well-known as occlusion problem. There are many techniques developed to solve occlusion and they can be divided into two groups [2]: Merge-Split (MS) approach and Straight-Through (ST) approach. In MS approach, when occlusion happens, the information about the occluded objects is frozen and new track is created for the group. When an occluded object leave the group its track is returned as it was before occlusion by measuring its similarity with the frozen tracks related to the group. In ST approach, when occlusion happens, the group is segmented to obtain each individual region of object. Compared to MS, this approach can achieve more accurate tracking during the occlusion but the drawback is that the computational cost is expensive. Moreover, in case the objects are fully occluded and the occlusion occurs for long duration, the track could be lost. Hence, in this study, MS approach is preferred for implementation.

The performance of MS techniques depends mainly on two things: merging-splitting event detection and similarity measure. Common approach for detecting merging and splitting is by using bounding box of an object [3-4]. When two or more bounding box is overlapped, occlusion seems to happen. When group of object splits into its members, similarity measure is performed to match the object based on its information before occlusion. Here, fuzzy similarity measure is incorporated to perform similarity measure between objects. First, the color, texture, and shape features of the object are extracted and transformed to fuzzy representation. To measure similarity between features, Fuzzy Feature Contrast Model (FFCM) [5] and its modified versions [6] are used. These models are developed from Tversky's similarity concept which is based on human perception about similarity. Fuzzy approach is chosen because it is similar to the way human perceives things and it can deal with uncertainty and imprecision.

The remaining part of this paper is structured as follows: In section 2 the tracking method is described, feature extraction feature representation is described in section 3, Section 4 examines Fuzzy Feature Contrast Model with its modified versions is explained in section 4 and In section 5 experimental results are discussed. Finally conclusions are derived in section 6 with further research.

2. TRACKING METHOD
In this section, the tracking method is briefly described together with the occlusion handling algorithm. The overall tracking system is depicted in Figure 1 below.

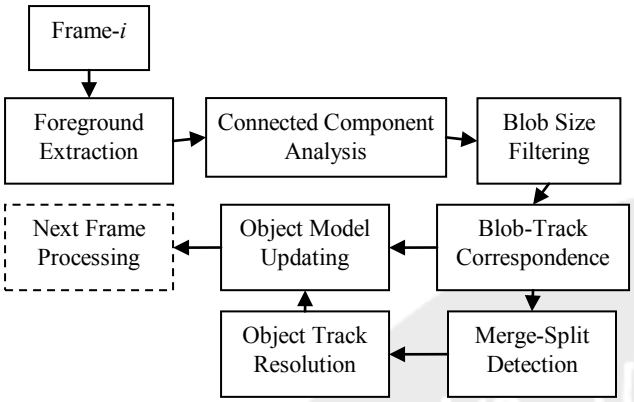


Figure 1. Structure of the tracking system

The first step is foreground extraction which attempts to acquire pixels belonging to the moving objects in the video. The resulting foreground pixels are then analyzed to search for connected components, so connected pixels are combined to form a structure which is called blob. Object correspondence between blobs and existing tracks is performed by using position and size information of the objects. Next, occlusion is handled separately by performing merging-splitting event detection and resolving occluded object tracks. Details about these processes are described in the following subsections.

2.1 Foreground Extraction

Foreground is the area in video which contains moving objects. In region based object tracking, the foreground should be separated from the background. This process is known as background subtraction. Various background subtraction methods have been developed in literature [7]. They differ in the way of modeling each pixel in the video frame. Among these methods, Adaptive Gaussian Mixture Models (GMM) [8] is one of which has better accuracy and moderately fast. Later, shadow removal is added to this method, so it can achieve better accuracy [9].

In adaptive GMM method, each pixel is modeled as a mixture of k gaussian distributions. This model is updated in time direction to adapt with changes in video sequence. When a particular pixel is examined in frame t, its value is matched to the model to decide whether the pixel belongs to background or not. The matching process takes into account the mean and standard deviation of each gaussian distribution. Shadow removal is performed by separating a pixel into its contrast and luminance components. If the chromatic and luminance value of the pixel are below some predetermined threshold, it is considered as shadow.

Connected component analysis is applied to the resulting foreground pixels to obtain groups of pixels connected each other which is called blob. After filtering out small size blobs, object regions are represented by the resulting blobs. Position and size of an object can be calculated directly by its blob, while more complex features such as color and texture are extracted from the frame by using the blob as mask.

2.2 Object Correspondence

Object correspondence is a core process in tracking. It maintains identity of an object in consecutive frames. In a particular time t, there exist m tracks which are maintained until frame t-1 and n blobs extracted from current frame t. The task is to establish correspondence between tracks and blobs. It is performed by calculating minimum distance between each pair of track and object. Let X, V, S, and ΔS denote position, velocity, size, and size difference of object an object respectively, the distance between track i and blob j is calculated as follows

$$d(i,j) = \left| X_i - \left(X_j + V_j \right) \right| + \left| S_i - \left(S_j + \Delta S_j / S_j \right) \right|, \quad \textbf{(1)}$$

where i and j refer to blob and track index. One-one correspondence is established by taking the pair (i,j) that yields minimum value for the above equation. Based on this result, any information related to each track is updated in the following way

$$T_i = \rho T_i + (1-\rho) T_{i-1}, \quad \textbf{(2)}$$

where T denotes any information related to the tracks, ρ is learning rate, indicates the extent to which the model adapt with new values.

Once correspondence established, one of the following conditions applies:

(1) all tracks and blobs are corresponded each other,

(2) there exists blob which does not correspond to any of the tracks, or

(3) there exists track which does not correspond to any of the blobs.

In case (2) there are two possibilities, either a new object entered the video or spliting has just happened. So, a check for new object and splitting event detection is called. In case (3), there are also two possibilities: either the object exited from video or occlusion happened. So, the check for object exiting and merging event detection is called. Merging and Splitting event detection are described in more details in the following subsection.

2.3 Merging and Splitting Detection

As stated in introduction, the performance of MS approach depends on merging and splitting detection beside similarity measure. So it is important to ensure that when objects merge or split, the tracking system can detect the events correctly. Here, the object bounding box is used to determine if objects merge and form a group. The merging detection procedure is illustrated in Figure 2 below.

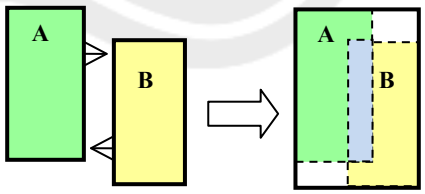


Figure 2. Merging event detection

If a track does not correspond to any one of the blobs, its predicted position is calculated and checked if the bounding box of the predicted object position is overlapped with any other object. The predicted bounding box is obtained by adding the previous bounding box by its recorded velocity as follows

$$\hat{B}_j^t = B_j^{t-1} + V_j. \quad \textbf{(3)}$$

Note that, the bounding box is defined as its four corner positions, so the above calculation is equivalent to matrix translation. Next, if a merging event is detected, occlusion handling is performed by firstly, the tracks related to occluded objects are frozen and secondly, new track is created for the group. Now, the group is treated as individual object. As long as the occlusion happens, the frozen tracks is kept and retrieved when the group splits.

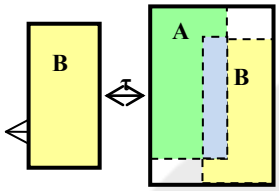


Figure 3. Splitting event detection

Splitting event is detected when a blob does not correspond to any existing track. If it is not a new object entered the scene, it is checked whether there is any occluded region around this blob in some distance not more than predetermined threshold. If it is true, it indicates that the object has just left the group. Next, occlusion is resolved by measuring the similarity between the splitted object and each frozen track related to this group. Here, fuzzy similarity measure (described in section 4) is incorporated to measure the objects similarity by their color, texture, and shape features. The resulting similarity value is used to determine the most probable track correspond to the object.

3. FUZZY FEATURE REPRESENTATION

Features are used as descriptor of an object or image so it can be distinguished with others. The more features used the more accurate object is represented. But too many features can cause what is called curse of dimensionality where increasing number of features reduce the significance of other features. This section examines features and their representations that are used in the proposed method.

3.1 Feature Extraction

In this study, features are needed for measuring the similarity between objects once they split from group. Here, color, texture, and shape features are extracted from the object. The color feature is represented as three-dimensional color histogram in RGB color space. As a trade-off between computational cost and accuracy, the number of bins in each channel is chosen as 16, so the total number of histogram equals to 16×16×16=4096. The histogram is calculated with masking, so only those pixels within object area are considered.

For representing texture feature, Local Binary Pattern [10] is used. For each pixel in object area, its difference is calculated with 8 neighbor pixels in certain radius by counter clock-wise circular scanning. The difference values are then binarized by using some threshold value. The resulting bit string is then converted to decimal number and quantized in form of histogram. Masking is also used to perform the more accurate feature extraction.

Shape is usually represented as complex feature extracted from image or object contour. However, for tracking system in real world scenes, objects belonging to the same type is typically difficult to distinguish by their shapes. For instance, human contour shape is very similar each other. Here, shape feature is more useful for differentiating objects with different types. Besides, extracting high level shape feature from countour requires high computational cost so it is not well-suited for real time application. In the proposed method, simple shape representation, aspect ratio and size of the object (in pixels) are used. These features had been used before in previous research [11].

For each object in the video, an appearance model is built from all of these features. During its existance, the feature is extracted and the corresponding model is updated.

3.2 Fuzzy Feature Representation

The advantage of fuzzy approach is that it can express uncertainty aspects of particular problem explicitly. Fuzzy allows partial truth values as human being does, so the problem is perceived in more intuitive way. In this study, extracted color, texture, and shape features are transformed into fuzzy representation.

As mentioned before, both color and texture are represented in form of histogram. There are several fuzzy representations developed for color histogram. The common approach is to divide histogram into several partitions and membership function is built for each partition. Other methods use simpler way to represent fuzzy histogram by dividing each bin by maximum bin value [12] or the sum of all bin values. Here, the later is used for fuzzy representation of RGB and LBP histogram. The resulting bin values are in range [0,1].

Aspect ratio and object size are represented by fuzzy predicates. The predicates are Low, Medium, and High for aspect ratio and Small, Medium, and Large for object size. For each predicate, a membership function is defined and the feature values is determined by evaluating these functions. Now, the shape feature consists of six values, one for each fuzzy predicates.

4. FUZZY SIMILARITY MEASURE

Fuzzy similarity measure has been used in region based image retrieval systems [6,13,14]. Here, the methods which are based on Fuzzy Feature Contrast Model are used. In [14], this model is successfully used for region based CBIR system. As the developed system is region based object tracking, the problem is somewhat similar.

First introduced in [5], FFCM is formulated based on Tversky's similarity concept. The main idea behind this model is that the more common features between two objects the more similar they are and the more different features between them the less similar they are. In set theory, the common and different features can be seen as set intersection and set difference. So, the idea is clearly shown as the equation below:

$$S(A,B) = f(A \cap B) - f(A - B) - f(B - A) \qquad \textbf{(4)}$$

Using fuzzy set theory, common set intersection is min(A,B) and set difference is defined in [5] as max(A-B,0). The original FFCM is as follows

$$S = X - (\alpha.Y + \beta.Z). \qquad \textbf{(5)}$$

where X, Y, and Z denote A∩B, A-B, and B-A respectively. α and β determine the asymmetry property of the model when $\alpha \neq \beta$. In [6], the model is modified while maintaining its property,

$$S_{Jaccard} = \frac{X}{X+Y+Z}, \quad \textbf{(1)}$$

$$S_{Dice} = \frac{2X}{2X+Y+Z}, \quad \textbf{(2)}$$

$$S_{Jaccard} = \frac{X}{X+Y+Z}. \quad \textbf{(3)}$$

Compared to the original FFCM, these models can avoid values less than zero when f(A-B)+f(B-A) is greater than f(A∩B). It is useful because interpreting negative similarity value is somewhat difficult. For histogram representation, the values X, Y, Z are obtained by performing operation for each pair of bin values and then summed. While for shape feature, the operation is performed on fuzzy predicate. These operations are expressed as follows

$$X = \sum_i \min(\mu(A_i), \mu(B_i)) \quad \textbf{(4)}$$

$$Y = \sum_i \max(\mu(A_i) - \mu(B_i, 0)) \quad \textbf{(5)}$$

$$Z = \sum_i \max(\mu(B_i) - \mu(A_i, 0)) \quad \textbf{(6)}$$

To obtain the best suited model for tracking problem domain, an experiment is conducted to compare their performance. Object images are cropped from video frames so that each object is represented by five images. A simple content based image retrieval is built where each image in database is used as query. Query accuracy as calculated as the number of images corresponding to the same object as query image in top five query result. The total accuracy is obtained by averaging all query accuracy. Note that the performance is not measured for shape features since in this method shape feature is used to distinguish type of objects not the object itself. The result of the experiment showed that the three modified versions perform better than original FFCM. Among these models, $S_{Jaccard}$ is considered more simple than the other two models. So, in the proposed method this model is used to perform similarity measure for color and texture feature. For shape feature, original FFCM model is used since the experiment in [5] showed that the model is effective to measure shape feature similarity. The total similarity is calculated as weighted average values over color, texture, and similarity values

$$S_{Total} = \alpha S_{Color} + \beta S_{Texture} + \gamma S_{Shape} \quad \textbf{(7)}$$

where α, β, and γ are weights for each similarity component and $\alpha + \beta + \gamma = 1$.

5. EXPERIMENT RESULTS

The experiment is conducted to show the effectiveness of the proposed method to handle occlusion in real world scenes. The data used in the experiment are obtained from a real traffic in Universitas Indonesia campus, Depok. The objects in video can be human, car, motor bike, or bicycle.

Figure 4 shows the result of the proposed method on real video scenes. There are two objects in video: the first one, marked by red box is a group of people walking together and the second one, marked by green box, is a motor bike. When the two objects occluded, merging event is detected and the occlusion is handled by creating new track, marked by blue box, regarding the newly formed group. Meanwhile, the track of individual object is frozen during occlusion. When the group splits, these tracks are recalled. The similarity measure is performed to restore the tracks as it was before occlusion happened.

5.1 Dealing with Inaccurate Blob Extraction

The second row of the image shows the corresponding extracted blobs in each frame. As shown in the figure, the extracted blobs are somewhat not accurate. In practice, it is difficult to obtain very accurate blob result in every situation. Background subtraction could perform poorly if the foreground is similar to the background and in presence of fast illumination change. Moreover, the method depends on many parameters to fit in certain condition. So, the challenge to the tracking system is to deal with this limitation.

Given this situation, the proposed method can still perform similarity measure effectively. It is because the features are represented as fuzzy and the similarity is measured over this representation. The update procedure can avoid instability of feature extraction caused by inaccurate blob extraction. Moreover, by combining color, shape, and texture the system can survive in hard conditions such as illumination change or when the objects are similar each other. The result shows the advantage of fuzzy approach in dealing with uncertainty and imprecision.

5.2 Computational Time Analysis

Running time of the proposed method can be analyzed by dividing the overall process into its components. The major processes are: background subtraction, connected component analysis, object correspondence, feature extraction and model updating, and occlusion handling. These processes run for each retrieved frame in sequential manner.

For background subtraction, Adaptive GMM technique requires $O(m)$ time for each pixel [7], where m is the number of Gaussian distribution. Connected component analysis is performed by using linear time algorithm proposed in [15], so the required time is $O(h \times w)$, where h and w are height and weight of the video frame and $h \times w$ equals to the total number of pixels in a frame. In object correspondence, each pair of blob and tracks are examined, so the total time is $O(k^2)$ where k is maximum number of objects. For feature extraction and model updating, the running time is determined by the number of bins in color histogram $(b_c)^3$ multiplied by the maximum number of objects k and the blob size s. Whilst for occlusion handling, the process is dominated by similarity measure for occlusion resolution. The complexity is determined by the number of bins for color histogram b_c and LBP histogram b_t. The overall complexity is the sum of all component which is $(h \times w \times m)+(h \times w)+(k^2)+(s \times k \times ((b_c)^3+b_t))+(k \times ((b_c)^3+b_t))$.

In the experiment, the system is run in 2 GHz Intel Pentium Dual CPU with 2 GB of RAM. The average processing time per frame was 94.5 milliseconds, so the frame processing rate is ≈ 10. This rate is fast enough for CCTV video surveillance which do not employ high frame rate for storage efficiency reason.

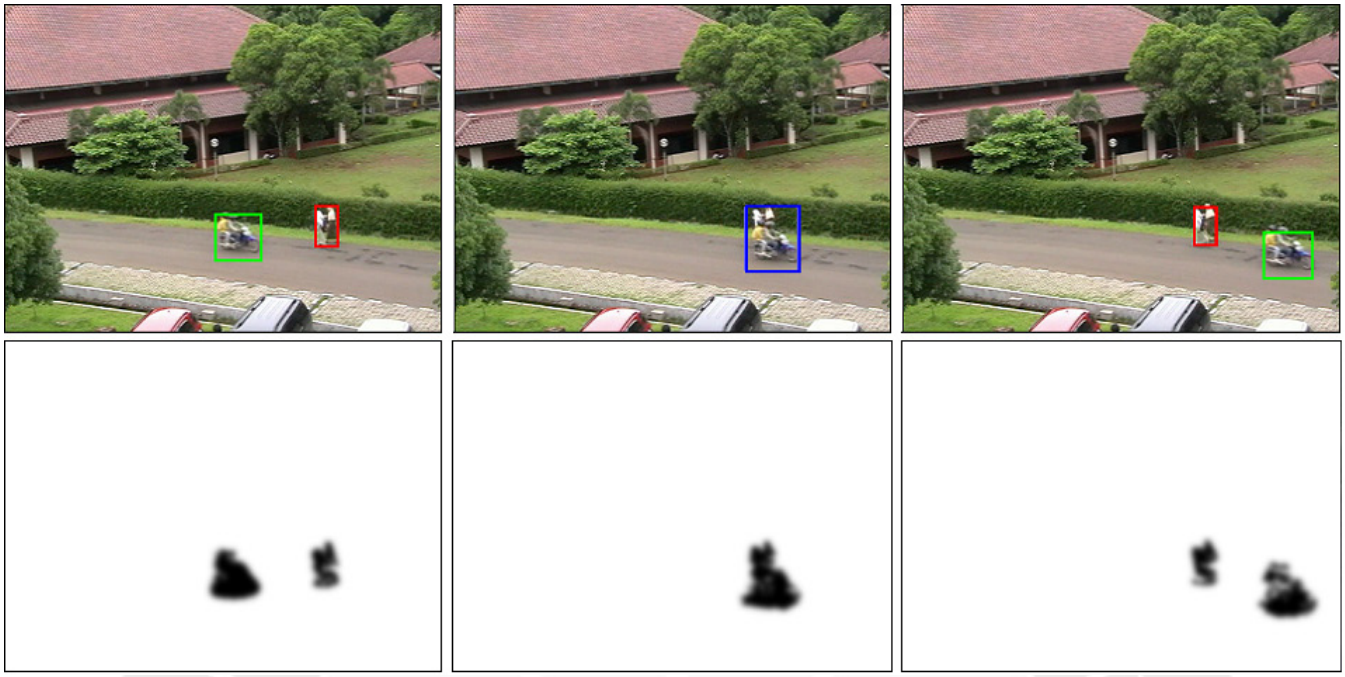


Figure 4. The tracking result in real traffic video sequence

6. CONCLUSIONS

Occlusion is a very common problem in multi-object tracking. It can degenerate the tracking accuracy if not properly handled. In this paper, a method for resolving occlusion is proposed. The occlusion handling procedure consists of two parts: merging-splitting detection and resolving object identity after occlusion. Merging event is detected by using object bounding box while splitting event is detected by searching any group around a split object. When an object leaves its group, the track is resolved by performing similarity measure to find its corresponding track before occlusion happened. Here, fuzzy similarity measure is incorporated by representing object by fuzzy feature and measuring its similarity using Fuzzy Feature Contrast Model and its derivative. The experimental results showed the effectiveness of the proposed method. The results also showed that the method runs in moderately fast time and hence, is suitable for real video surveillance application.

7. REFERENCES

[1] Hu, W., Tan, T., Wang, L., Maybank, S. 2004. A Survey on Visual Surveillance of Object Motion and Behaviors. IEEE Trans On System, Man and Cybernetics – Part C: Applications and Reviews, 34(3).

[2] Gabriel, P. F., Verly, J.G., Piater J. H., Genon, A. 2003. The State of the Art in Multiple Object Tracking Under Occlusion in Video Sequences. Proc. of Adv. Concepts for Intelligent Vision Syst. (ACIVS, Sep. 2003): 166-173.

[3] Javed, O., Shah, M. 2002. Tracking and Object Classification for Automated Surveillance. LNCS; Proc. of the 7th European Conf. on Computer Vision-Part IV, 2353: 343-357, Springer-Verlag.

[4] Yang, T., Li, S. Z., Pan, Q., Li, J. 2005. Real-time Multiple Objects Tracking with Occlusion Handling in Dynamic Scenes. Proc. CVPR'05 – 1: 970-975.

[5] Santini, S. & Jain, R. 1999. Similarity Measures. IEEE Trans. *PAMI*, 21(9): 871-883, Sep 1999.

[6] Omhover J. F., Detyniecki, M., Bouchon-Meunier, B. 2004. A Region-Based Image Retrieval System. Proc. of IPMU'04, Perugia, Italy, July 2004.

[7] Piccardi, M. (2004). Background Subtraction Techniques: a Review. IEEE Intl. Conf. on Systems, Man and Cybernetics (Oct 2004), 4: 3099-3104.

[8] Stauffer, C. & Grimson, W. E. L. 2002. Adaptive Background Mixture Models for Real-Time Tracking. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (2), 1999.

[9] KaewTraKulPong P. & Bowden, R. 2001. An improved adaptive background mixture model for real-time tracking with shadow detection. Proc. 2nd EU Workshop on Adv. Video-Based Surveillance Syst, Sep 2001.

[10] Mäenpää, T. & Pietikäinen, M. 2005. Texture Analysis with Local Binary Pattern. Ch 1, in C. Chen and P. Wang (eds) Handbook of Pattern Recognition and Computer Vision, 3rd ed: 197-216. World Scientific.

[11] Mardiko, R. & Widyanto, M. R. 2009. An Approach to Object Counting for Video Surveillance Using Fuzzy Inference System with Fault Tolerance. Proc. Intl Conf on IT Application and Management, UI Depok, Apr 2009.

[12] Van der Weken, D., Nachtegael, M., and Kerre E. 2003. Using Similarity Measures for Histogram Comparison. In: De Baets,

B., Kaynak, O., Bilgic, T. (eds) IFSA 2003, LNCS, 2715: 396-403. Springer-Verlag Berlin Heidelberg.

[13] Chen, Y. X., Wang, J. Z. 2002 . A Region-Based Fuzzy Feature Matching Approach for Content-Based Image Retrieval. IEEE Trans. PAMI, Vol 24 No. 9, September 2002

[14] Jiang, W., Er, G., Dai, Q., Gu, J. 2006. Similarity Based Online Feature Selection in Content Based Image Retrieval. IEEE Trans. on Image Processing, Vol. 15(3), March 2006.

[15] Chang, F. 2004. A linear-time component-labeling algorithm using contour tracing technique. Computer Vision and Image Understanding, 93(2):206-220.

Search Engine Application Using Fuzzy Relation Method for E-Journal of Informatics Department Petra Christian University

Leo Willyanto Santoso
Petra Christian University
Siwalankerto 121-131
Surabaya 60236 INDONESIA
+6231-2983455
leow@petra.ac.id

Rolly Intan
Petra Christian University
Siwalankerto 121-131
Surabaya 60236 INDONESIA
+6231-2983455
rintan@petra.ac.id

Prayogo Probo Susanto
Petra Christian University
Siwalankerto 121-131
Surabaya 60236 INDONESIA
+6231-2983455

ABSTRACT
Nowadays, scientific articles are easily obtained, because many researchers who conduct research discover new things. However, the increasing number scientific articles is not accompanied by the availability of applications to assist in the search for relevant articles. Today, available search engine applications perform only a search process based on string matching of search terms. In this research, a search engine application based on keyword relevance by using fuzzy relationship was developed.

This search engine application is built using PHP programming language and mysql as its database. Windows XP is used as the operating system. The used methods in fuzzy relationship are keyword used to paper, paper to paper, and paper to keywords and keyword to keyword. In addition, the components used to convert pdf files into plain text format.

Based on the results of experiments conducted, the process of searching for the 25 articles takes less than 5 seconds. For the indexing process, it is influenced by the number of pages per article.

Keywords
Fuzzy Relation, Search Engine, Paper.

1. INTRODUCTION
At this time, journal is one of the many forms of documents selected by the researchers and scientists to put the results of experiments or research that have been conducted. Through the journal, the researchers poured all aspects of the research that been conducted by attaching a detailed information about the research he had conducted.

A journal as a medium of information from the experts/researchers to the public media has an important role and very strategic. For example in the field of education, the journal serves as a good material for teaching materials (for teachers) or as a reference for students to learn a new science.

Currently, website has been highly developed. This resulted in a website used as a medium of publication of a journal from experts/researchers to the public or known by the name of the e-journal. But more and more of the e-journal are not followed by the use of search engines technology. The search engine on each of the e-journal that is useful to facilitate a user who wants to do a search on a journal and other journal/articles that may still relate to one another is needed.

The problem is how to design search engines on e-journal that can produce a related mutually journal to one another based on the keywords that are input by the user.

This paper presents a new search engine applications that do not only search on the similarity keyword provided by journal or scientific paper, but also provide a reference paper which relate to each other as well as journal is desired by the user.

The remaining part of this paper is organized as follows. Section 2 presents an overview of current proposal for dealing with fuzzy relation. Section 3 depicts the approach that we have delineated to solve the proposed problems. Section 4 discusses the performance of proposed methods. Finally, section 5 concludes the paper.

2. FUZZY RELATION
Fuzzy relation is a method for explaining the relationship of two different things (completely different). As illustration, the word "apple" (apple) and "tiger" (tiger) then in general the two words are not related. In general, the word "apple" refers to the name of the fruit and the "tiger" refers to the name of wild animal.

In the computer world there is manufacturer software, Macintosh (Mac). Mac has the brand "apple" so often referred to as the Apple Macintosh. Recently, Mac issued a new operating system called "Tiger" OS. From the relationship with Mac as the word "Apple" and "Tiger" is actually not related in general and in writing, have a relationship in the world of computers. Given the fuzzy relation then this kind of relationship will be examined with an assumption and goal that by knowing the relationship closeness/kinship between the two word/object. In relation to the world of search (searching), then by inserting the word "apple", there is the possibility of the word tiger will also be a result of output. Not because the results wrong, but because between the word "apple" and "tiger" there is kinship [4].

Explanation of fuzzy relation can also be described as follows: two words that completely unrelated (eg: "apple" and "tiger"), will have a relationship when both the word is addressed in one document. More and more documents that discuss both the relationship between the two words ("apple" and "tiger") will be getting closer.

Fuzzy Relation will search 4 links from a combination of words (keywords) and documents (paper) these relationships are:

- *Keyword* to *paper*
- *Paper* to *paper*
- *Paper* to *keyword*
- *Keyword* to *keyword*

Explanation of each of this relationship along with the calculation process is described as follows:

1. At this step is assumed relationship between the keyword to the weight of the paper has value to the paper of the following keywords:

   $P = \{P_1, P_2, \ldots Pn\}$ is a *set of papers*

   $D = \{D_1, D_2, \ldots Dn\}$ is a *set of keywords*

   For example from the data obtained by paper and keyword relationship expressed as a fuzzy set of papers on the following keywords:

   $P_1 = \{0.3/D_2, 0.7/D_5, 1/D_7, 1/D_8\}$,

   $P_2 = \{1/D_2, 0.8/D_5, 0.8/D_7, 1/D_8\}$,

   $P_3 = \{0.9/D_1, 0.9/D_3, 1/D_4, 0.8/D_6\}$,

   $P_4 = \{1/D_1, 0.5/D_3, 0.8/D_4, 0.8/D_6\}$,

   $P_5 = \{0.1/D_2, 0.7/D_5, 1/D_4, 1/D_8\}$,

   $P_6 = \{0.9/D_2, 1/D_5, 0.8/D_4, 1/D_8\}$

   For $P = \{P_1, P_2, \ldots P_6\}$ and $D = \{D_1, D_2, \ldots D_8\}$, where each *paper*/document regarded as a fuzzy set of keywords so we get $\mu P_5\ (D_1) = 0.1$.

2. *Similarity* between 2 *papers* expressed as a function of R where $R: P \times P \rightarrow [0.1]$ [5]

$$R(P_i,P_j) = \frac{\sum_D (\mu_{Pi}(D), \mu_{Pj}(D))}{\sum_D \mu_{Pj}(D)} \qquad (2.1)$$

   Where:

   R: Relation

   P_i: *Paper*/document i

   P_j:*Paper*/document j

   D: *Keyword*

   μ: *Membership function* as a *mapping* $\mu P_i: D \rightarrow [0,1]$.

   Can find a relationship between a paper with one another, eg:

$$\text{relation } R(P_1,P_2) = \frac{0.3+0.7+0.8+1}{1+0.8+0.8+1} = \frac{2.9}{3.6} = 0.78$$

   The calculation of paper to paper as a whole can be seen in Table 1.

Table 1. Relation *Paper* to *Paper*

X / Y	P1	P2	P3	P4	P5	P6
P1	1,00	0,78	0	0	0,64	0,54
P2	0,93	1,00	0	0	0,64	0,73
P3	0	0	1,00	0,97	0,36	0,22
P4	0	0	0,83	1,00	0,29	0,22
P5	0,60	0,50	0,28	0,26	1,00	0,70
P6	0,67	0,75	0,22	0,26	0,93	1,00

3. From the existing data of keywords related to the paper, then we will get the paper on the relationship between keywords. Relationship of paper to keyword can be calculated using formula 2.2 [5]:

$$H_{Dj}(P_j) = \frac{\mu_{pi}(D_j)}{\mu_{pi}(D_1) + \mu_{pi}(D_2) + \ldots + \mu_{pi}(D_m)} \qquad (2.2)$$

   Where:

   R: Relation

   P_i: *Paper*/document i

   P_j: *Paper*/document j

   D: *Keyword*

   μ: *Membership function* as a *mapping* $\mu P_i: D \rightarrow [0,1]$.

   Example: Calculate the weight of *keyword* (d2) for *paper* 1

$$\eta_{Dj}(P_j) = \frac{0.3}{0.3+0.7+1+1} = \frac{0.3}{3} = 0.1$$

   so the final result is:

   $D_1 = (0.25/P_3, 0.32/P_4)$,

   $D_2 = (0.1/P_1, 0.28/P_2, 0.06/P_5, 0.24/P_6)$,

   $D_3 = (0.25/P_3, 0.16/P_4)$,

   $D_4 = (0.28/P_3, 0.26/P_4, 0.36/P_5, 0.27/P_6)$,

   $D_5 = (0.23/P_1, 0.22/P_2, 0.25/P_5, 0.27/P_6)$,

   $D_6 = (0.22/P_3, 0.26/P_4)$,

   $D_7 = (0.33/P_1, 0.22/P_2)$,

   $D_8 = (0.33/P_1, 0.28/P_2, 0.36/P_5, 0.27/P_6)$

4. *Similarity* between 2 *keywords* expressed as a function of R where $R: D \times D \rightarrow [0,1]$ as written in the formula 2.3 [5]:

$$R(D_i,D_j) = \frac{\sum_P \min(\eta_{D_i}(p), \eta_{D_j}(p))}{\sum_P (\eta_{D_i}(p))} \qquad (2.3)$$

   Where:

   R: Relation

   P_i: *Paper*/document i

   P_j: *Paper*/document j

   D: *Keyword*

μ: *Membership function* as a *mapping* $\mu P_i: P \rightarrow [0,1]$.

can be found the relationship between keywords with each other eg.:

$$\text{relation } R(D_1,D_3) = \frac{0.25+0.16}{0.25+0.16} = 1$$

$$\text{relation } R(D_3,D_1) = \frac{0.25+0.16}{0.25+0.32} = 0{,}72$$

Calculation of *keyword to keyword* as a whole can be seen in Table 2.

Table 2. Relationship between *Keyword* to *Keyword*

X / Y	D1	D 2	D 3	D 4	D 5	D 6	D 7	D 8
D 1	1,00	0	1,00	0,44	0	1,00	0	0
D 2	0	1,00	0	0,26	0,64	0	0,58	0,53
D 3	0,72	0	1,00	0,35	0	0,79	0	0
D 4	0,89	0,44	1,00	1,00	0,54	1,00	0	0,51
D 5	0	0,91	0	0,44	1,00	0	0,82	0,86
D 6	0,84	0	0,93	0,41	0	1,00	0	0
D 7	0	0,47	0	0	0,46	0	1,00	0,49
D 8	0	1,00	0	0,54	1,00	0	1,00	1,00

.

3. SEARCH ENGINE APPLICATION

Chapter 3 describes the design of systems that are the basis of the developing the application of scientific journal search engines in this research. Basically there are two main processes in this application; they are the indexing process and searching process.
Indexing process takes the longest runtime when compared with the process of searching. It is caused by a cutting process of a document into a word. Indexing process execution depicted in Figure 1.

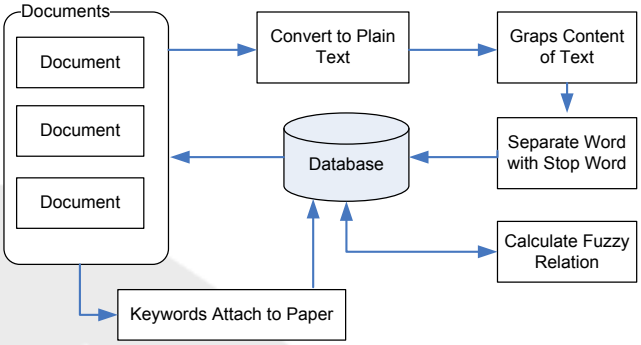


Figure 1. Indexing process

Here is a sequence of description of the indexing process:

1. Retrieve a document that has no plain text file / not yet indexed.
2. Convert the pdf documents into plain text files that are readable by the programming language PHP 5.
3. Input a plain text file data into the database to be stored and through the next process.
4. Read the contents of plain text files and store a long string into an array.
5. Enter a keyword from a document originating from the author to serve as the keyword / main keywords on applications and enter into a database. Keyword is better known by the name of keywords attached to paper.
6. Grab a few sentences from the abstract and titles for snipset when displaying search results on the process of searching.
7. Take the author's name and affiliation to be used in the process of searching with the author and affiliation search mode.
8. Rupture long string that is stored in an array into a word-per-word and separate from the existing stop word. Next save all the words that have been cut into the database and delete its stop word.
9. The last process of this indexing process is the process of calculating the value of fuzzy relations.

- Searching Process

The search engine used 2 kinds of fuzzy methods to process the searching process, both methods are:

a. Ordinary Fuzzy Method

Ordinary fuzzy method is used when a user input keyword has never been through the process of indexing, but the keywords are included on one or more papers that are stored on the system.

In this method, the fuzzy value calculation is only carried out on the basis of number of word occurrence; it means the fuzzy value calculation is performed only for papers that have keyword input on its content. This method involves only one fuzzy process has been done before in the indexing process, it is the calculation of keywords to paper fuzzy.

b. Extended Fuzzy Method

Unlike the ordinary fuzzy method, the calculation by this method not only involves the number of word occurrence alone but also involves all fuzzy calculations have been

done on indexing process. With this method, papers that do not have keyword input on its content will be found.

The use of two methods is for the speed/runtime purpose when the searching process is running. In this search engine, users can not choose the method that will be used, because the system will automatically check the incoming keyword to further select one of two methods for implementing the process of searching. General description of the process of searching can be seen in Figure 2.

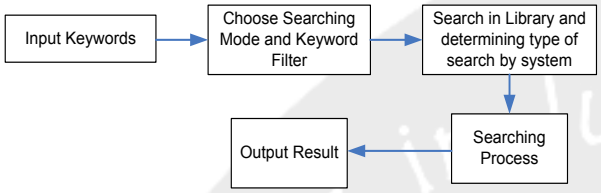


Figure 2. Searching process

4. EXPERIMENTS

In this section, we present an experimental result of new search engine application. This system was built in PHP [1, 2, 3] on a PC with 2.4 GHz Pentium ® 4 CPU and 1 GB of RAM under MS Windows XP Pro.

4.1 Searching Type - *Extended Fuzzy*

Search by extended fuzzy type is a searching type which looks for related papers of related keywords that are input by the user. The implementation of an extended type of fuzzy search with keyword input - one word can be seen in Figure 3.

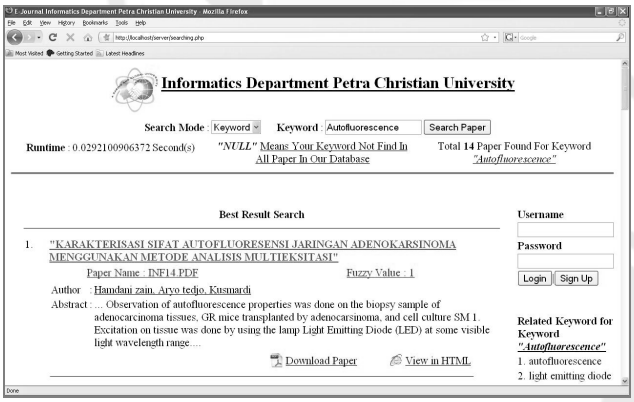


Figure 3. Extended type of fuzzy search with keyword input - one word

The implementation of an extended type of fuzzy search with a keyword input more than one word/phrase can be seen in Figure 4.

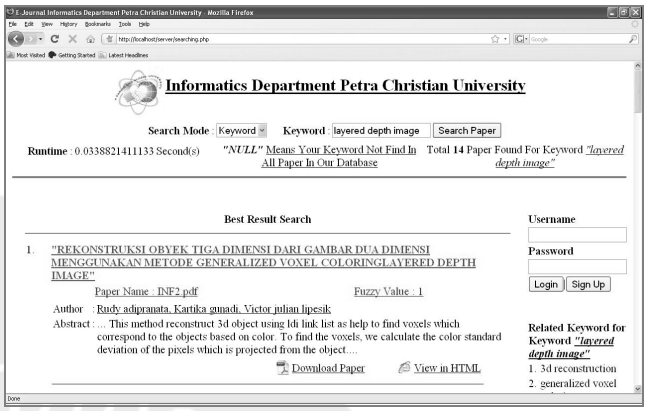


Figure 4. Extended type of fuzzy search with keyword input – more than one word/phrase

4.2 Searching Type - *Ordinary Fuzzy*

Searching by ordinary fuzzy type is a search that only looks for keywords in the paper, inputted by the user. Ordinary type of fuzzy search implementation with a single word keyword input can be seen in Figure 5.

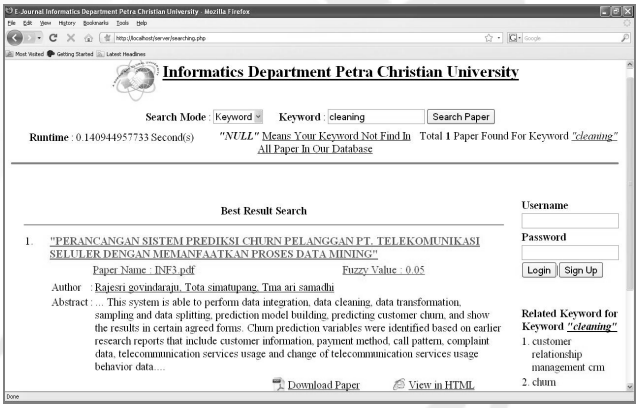


Figure 5. Ordinary type of fuzzy search with a single word keyword input

Ordinary type of fuzzy search implementation with the input of more than one word keywords/phrases can be seen in Figure 6.

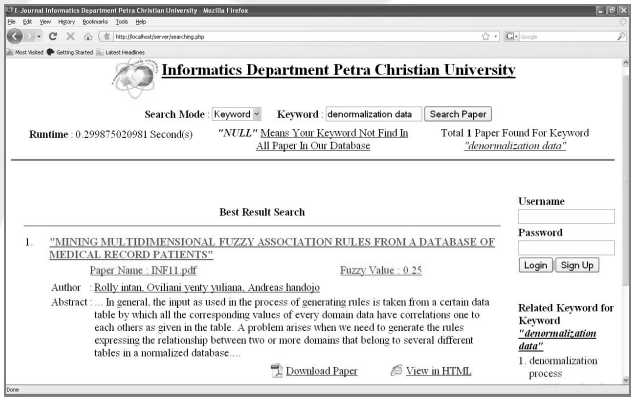


Figure 6. Ordinary type of fuzzy search with more than one work/phrase keyword input

4.3 Search involving symbol and *Stop Word*

In addition to testing with the extended fuzzy search types and ordinary fuzzy, involving a stop word and symbol are also important. This is related to all rules that apply to the program. All these rules are based on assumptions that are not researched before, but just based on observations from the passage of searching module developing. In addition, this test is also a test of all rules that apply to the application.

The Implementation of the search involving symbol and stop word with one-word keywords can be seen in Figure 7.

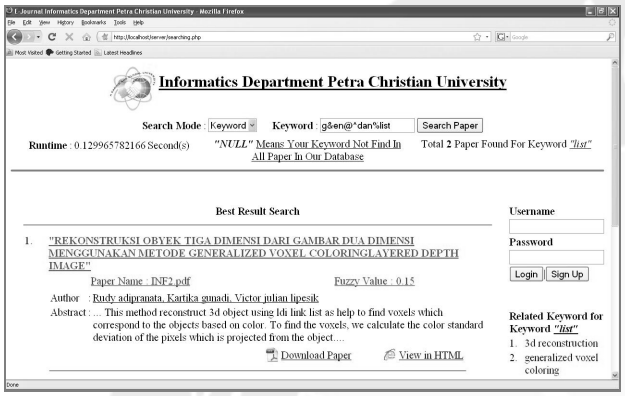


Figure 7. Search involve symbol and stop word with one word keyword input

Search engine implementation which involves symbols and stop word to the keyword input of more than one word/phrases can be seen in Figure 8.

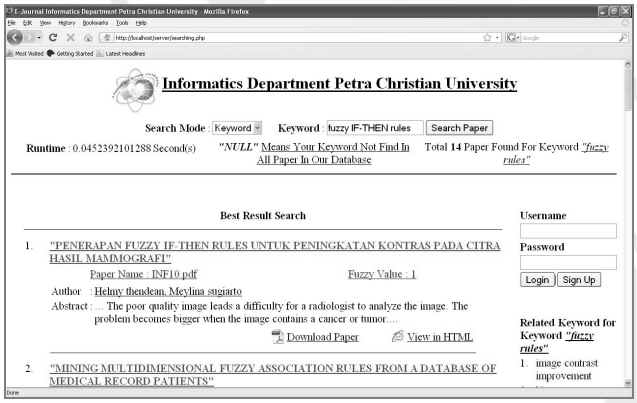


Figure 8. Search engine which involves symbol and *stop word* with more than one word/phrase keyword input

4.4 Searching involves *Keyword Attach to Paper*

Searching involving keyword attached to paper is a searching process performed on the input found on the library tables. This Search use extended fuzzy type because the keyword input is found in the library table. Search engine implementation involves keyword attached to paper with one word keyword input can be seen in Figure 9.

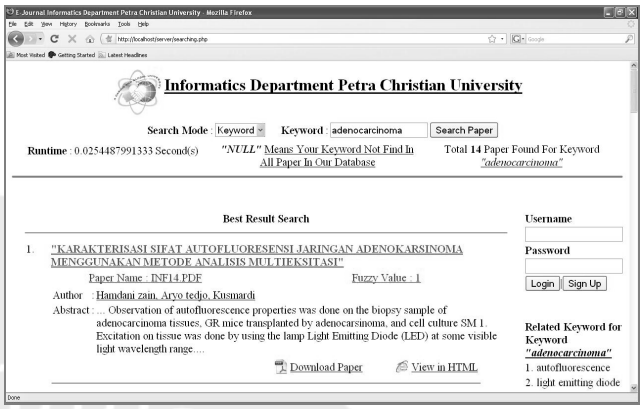


Figure 9. Search engine which involves *keyword attach to paper* with one keyword input

Search engine implementation which involves keyword attach to paper with more than one word/phrases input can be seen in Figure 10.

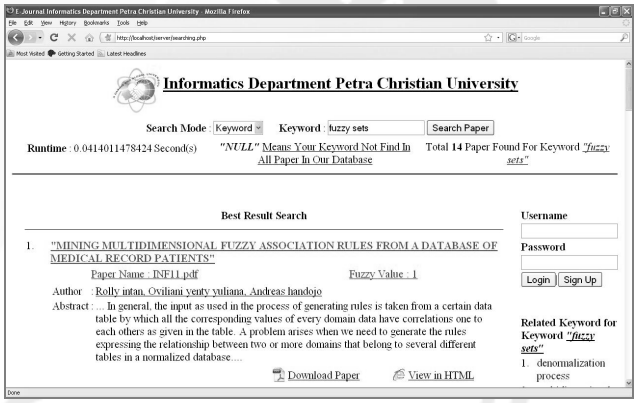


Figure 10. Search engine which involves *keyword attach to paper* with more than one word/phrase keyword input

4.5 Runtime Process

From the experiments conducted, it can be calculated an average speed of the process of searching on search engine applications. The average results of the calculation process of some kind of keyword search can be seen in Table 3.

Table 3. Runtime test result

User inputed Keyword	***Runtime* (second)**
Autofluorescence	0.0254549980
layered depth image	0.0606830120087
Data	0.0613670349121
denormalization *data*	0.343075037003
g&en@^e$rate	0.105587005615
fuzzy IF-THEN rules	0.261485099792

adenocarcinoma	0.0284929275513
fuzzy sets	0.0446968078613
The Average of Runtime	0.8831452369523/8 = **0.1103931546190375**

From 8 keywords input in the system, the average of searching process is 0.1103931546190375 seconds.

Runtime test of indexing process is the main focus of all the testing and experiments performed on the system. From several experiments, it is resulted that more and more number of papers and the number of keywords, the runtime required to complete the indexing process is also getting bigger. From the seventh step in the indexing process, step-core indexing and fuzzy relationship value calculation takes the greatest runtime. The example of execution indexing process is presented in Figure 11.

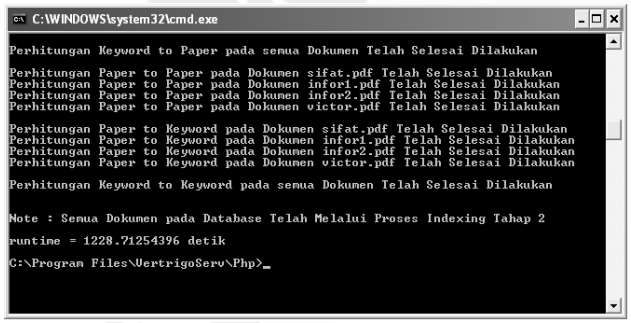


Figure 11. Indexing process

From the testing process it can be described that it takes 1228.71254396 seconds or approximately 20 minutes 5 seconds to index four papers.

5. CONCLUSION

This paper deals with the implementation of fuzzy relation method to support e-journal search engine. Fuzzy value ranking system should be implemented into the calculation of fuzzy relations with respect to the accuracy of produced output.

The Implementation of search engine applications on operating systems other than Windows XP can be done by changing some parts of the segment of the program and replace the use of customized components with the operating system used.

The emphasis of this paper was on feasibility – identification of possible approaches and development of methods to put them into practices.

We are currently working on the implementation process of indexing and searching in a lot of server/multi server. Next, concerns is the quality of the result.

6. REFERENCES

[1] Bibeault, B. and Yehuda, K. 2008. *JQuery in Action*. Manning Publications. USA.

[2] Boronczyk, T., Naramore, E., Gerner, J., Le Scouarnec, Y., and Stolz, J. 2009. *Beginning PHP 6, Apache, and MySQL 6 Web Development*. John Wiley and Sons.

[3] Castagnetto, J., Rawat, H., Schumann, S., Scollo, C., and Veliath, D. 2000. Professional PHP Programming. 3rd Ed. Wrox Press Ltd. USA.

[4] Darmadi, B.A. 2005. *Perancangan dan pembuatan search engine paper/karya ilmiah berbasis web dengan metode fuzzy relation.* Unpublished undergrade thesis, Universitas Kristen Petra, Surabaya.

[5] Intan, R and Masao, M. Toward a fuzzy Thesaurus Based on Similarity in Fuzzy Covering. Australian Journal of Intelligent Information Processing, vol. 8 no. 3, 2004. pp. 132-139.

The Use of Gabor Filter and Back-Propagation Neural Network for The Automobile Types Recognition

Gregorius Satia Budhi
Informatics Department
Petra Chistian University
Siwalankerto 121-131
Surabaya 60236, Indonesia
(62-31) 2983455
greg@petra.ac.id

Rudy Adipranata
Informatics Department
Petra Christian University
Siwalankerto 121-131
Surabaya 60236, Indonesia
(62-31) 2983455
rudya@petra.ac.id

Fransisco Jimmy Hartono
Informatics Department
Petra Christian University
Siwalankerto 121-131
Surabaya 60236, Indonesia
(62-31) 2983455
fr8jimmy@yahoo.com

ABSTRACT
Type of automobile is the general factor that makes automobile different from each other. But conventional sensor cannot detect the automobile and recognize its type. Because of those reasons, we made an experiment application that can count the number and recognize automobile based on its type. This application uses gabor filter for feature extraction and back-propagation neural network for training and recognizing type of automobile. The experiment was done using various parameters for back-propagation neural network and gabor filters. The experimental result shows that the best error rate of recognition results is 16%. It's done with the brightness condition not too low or high.

Keywords
Gabor Filter, Back-propagation Neural Network, Automobile Type Recognition.

1. INTRODUCTION
Along with technological advances, nowadays computer science relating to digital image processing and computer vision also has been implemented to help human's work. Therefore in this experiment, we developed an application which can calculate the number of cars and identify them based on their types. Input of this application is a video of the car entering parking area and being taken by using a video camera / webcam. The beginning process of this application is separating car objects from the other objects that have been captured by the video camera / webcam. Furthermore, this application will count the number of cars that has been successfully detected and recognize its type. The number of cars passing by and the type of the cars will be the output from this application.

This research was inspired from another research that has been done by our colleagues majoring in electrical engineering from Petra Christian University in 2001 [9]. On that research, they have examined the same topic, namely Vehicle Type Recognition. On that research, a recognition process is done by using gabor filter and template matching method. In our research, we change the method of recognition using back-propagation neural network, and reevaluated the best parameter / configuration of gabor filters.

2. IMAGE PROCESSING
2.1 Grayscale
Grayscale is the one technique in digital image processing to flatten pixel values of the three RGB values into one same value. This technique can be done using YUV color system, by converting RGB to YUV and then take the Y component (illumination) [1]. This is done by using equation 1.

$$gray = Y = 0.2989 \times R + 0.5870 \times G + 0.1140 \times R \quad (1)$$

2.2 Low Pass filter
Low pass filter is used to give the smoothing (blurring) effects in the image. Besides, low pass filter also can be used to eliminate noise in the image [8]. The template 3x3 of low pass filter is shown at Formula (2) [4].

$$template = \frac{1}{9}\begin{bmatrix}1 & 1 & 1\\ 1 & 1 & 1\\ 1 & 1 & 1\end{bmatrix} \quad (2)$$

2.3 Background Subtraction
Background subtraction can be used to identify a motion between two images or more. In order to use the background subtraction technique, all images must be capture in time as close as possible without changes in light conditions and use a fixed background. By doing the background subtraction then a moving object can be detected [2].

$$g(x,y)=\begin{cases} f(x,y) & \text{if } |f(x,y)-b(x,y)| > threshold \\ 0 & \text{if otherwise} \end{cases} \quad (3)$$

Descriptions:
g(x, y) = Result of pixel's value
f(x, y) = Foreground pixel's value
b(x, y) = Background pixel's value

2.4 Median filter
Median filter is used to reduce random noise without give any blurring effect at the line of the image like low pass filter. The process that occurs in the median filter is as follows, for each pixel value in several areas that contained in the template on the image will be sorted first, and then will be searched for its median value.

This median value is become the gray level value for the result image [4].

2.5 Gabor Filter

Gabor filter is made under the concept of the mammalian cortical simple cells [11]. By using gabor filter each point on the image will be processed to obtain its vector feature. 2-D gabor filter is harmonic oscillator that obtained by modulating 2-D sine wave at a particular frequency and orientation with a Gaussian envelope.

Gabor filter can be used to capture frequency information. By adjusting gabor filter at the specific frequency and direction, then the information of local frequency and orientation can be obtained [3]. Gabor filter can be divided into two different equations, one for describing the real part and the other to describe the imaginary part of gabor filter [10].

$$\Psi_r(x,y,w_0,\theta) = \frac{1}{2\pi\sigma^2} exp\left\{-\left(\frac{x'^2+y'^2}{\sigma^2}\right)\right\} \times \left[\cos w_0 x' - e^{-w_0\frac{\sigma^2}{2}}\right] \quad (4)$$

$$\Psi_i(x,y,w_0,\theta) = \frac{1}{2\pi\sigma^2} exp\left\{-\left(\frac{x'^2+y'^2}{\sigma^2}\right)\right\} \times \sin w_0 x' \quad (5)$$

where:

$$x' = x\cos\theta + y\sin\theta \quad (6)$$

$$y' = -x\sin\theta + y\cos\theta \quad (7)$$

$$\theta_m = \frac{\pi}{8} m \quad (8)$$

$$w_n = \frac{\pi}{2(2)^{\frac{n}{2}}} \quad (9)$$

Descriptions:
x, y : Pixel position in spatial domain
w_0 : Radial middle frequency
θ : Gabor rotation
σ : Gaussian deviation standard of x and y

3. NEURAL NETWORK

3.1 Multilayer Neural Networks

A multilayer neural network is a feed forward neural network with one or more hidden layers. Typically, the network consists of an input layer of source neurons, at least one middle layer or hidden layer of computational neurons, and an output layer of computational neurons [7].

The input layer accepts input signals from the outside world and redistributes these signals to all neurons in the hidden layer. The output layer accepts output signals, or in other words a stimulus pattern, from the hidden layer and establishes the output pattern of the entire network. Neurons in the hidden layer detect the features; the weights of the neurons represent the features hidden in the input patterns. These features are then used by the output layer in determining the output pattern [7]. A multilayer neural network with two hidden layers is shown in Figure 1.

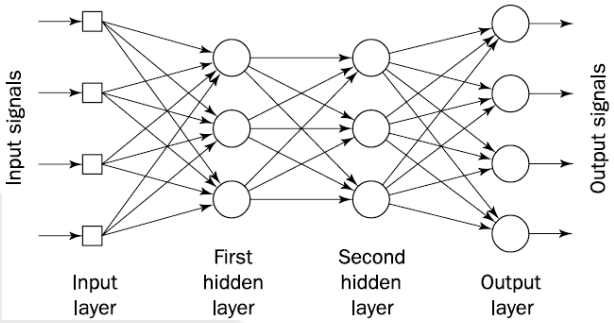


Figure 1. Multilayer neural network with 2 hidden layers [7]

3.2 Back-propagation Learning Algorithm

Step 1: Initialization

Set all the weights and threshold levels of the network to random numbers uniformly distributed inside a small range [5, 7]:

$$\left(-\frac{2.4}{F_i}, +\frac{2.4}{F_i}\right) \quad (10)$$

where F_i is the total number of inputs of neuron i in the network. The weight initialization is done on a neuron-by-neuron basis.

Step 2: Activation

Activate the back-propagation neural network by applying inputs $x_1(p), x2(p), \ldots, x_n(p)$ and desired outputs $y_{d,1}(p); y_{d,2}(p), \ldots, y_{d,n}(p)$.

(a) Calculate the actual outputs of the neurons in the hidden layer:

$$y_j(p) = sigmoid\left[\sum_{i=1}^{n} x_i(p) \times w_{ij}(p) - \theta_j\right] \quad (11)$$

where n is the number of inputs of neuron j in the hidden layer, and sigmoid is the sigmoid activation function.

(b) Calculate the actual outputs of the neurons in the output layer:

$$y_k(p) = sigmoid\left[\sum_{j=1}^{m} x_{jk}(p) \times w_{jk}(p) - \theta_k\right] \quad (12)$$

where m is the number of inputs of neuron k in the output layer.

Step 3: Weight training

Update the weights in the back-propagation network propagating backward the errors associated with output neurons.

(a) Calculate the error gradient for the neurons in the output layer:

$$\delta_k(p) = y_k(p) \times [1 - y_k(p)] \times e_k(p) \quad (13)$$

$$e_k(p) = y_{d,k}(p) - y_k(p) \quad (14)$$

Calculate the weight corrections:

$$\Delta w_{jk}(p) = \alpha \times y_j(p) \times \delta_k(p) \quad (15)$$

Update the weights at the output neurons:

$$w_{jk}(p+1) = w_{jk}(p) + \Delta w_{jk}(p) \quad (16)$$

(b) Calculate the error gradient for the neurons in the hidden layer:

$$\delta_j(p) = y_j(p) \times [1 - y_j(p)] \times \sum_{k=1}^{l} \delta_k(p) \times w_{jk}(p) \quad (17)$$

Calculate the weight corrections:

$$\Delta w_{ij}(p) = \alpha \times x_i(p) \times \delta_j(p) \quad (18)$$

Update the weights at the hidden neurons:

$$w_{ij}(p+1) = w_{ij}(p) + \Delta w_{ij}(p) \quad (19)$$

Step 4: Iteration

Increase iteration p by one, go back to Step 2 and repeat the process until the selected error criterion is satisfied.

4. APPLICATION DESIGN

This application is designed to be able to receive a stream image input from a video camera / webcam device in real time. The design of this application is shown in Figure 2.

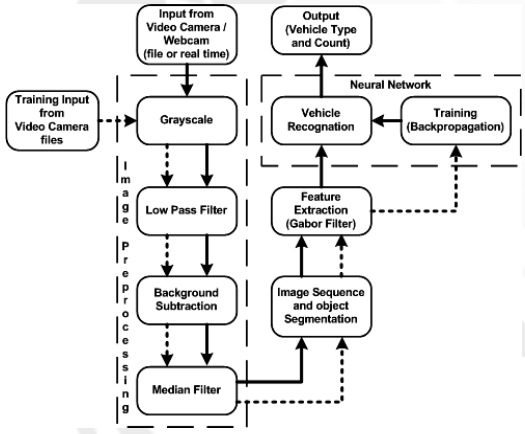


Figure 2. The application design.

4.1 Image Preprocessing Phase

In this phase, all the frames from video files or data streams will be processed using grayscale, low pass filter to eliminate noise, background subtraction to detect a moving object and the last process is the median filter to eliminate noise that may still remain. The example results of the preprocessing phase are shown in Figure 3.

Figure 3. Examples of image preprocessing phase result.

4.2 Segmentation Phase

The next phase is the segmentation phase. At this phase, each image frame in the video file (data stream) will be checked first on the most left, most right and middle vertically. If not all pixel in the middle position is black, while the most left and right area is black, it can be assumed that the frame in the checked area contained a complete car object (not truncated). Furthermore, in the frame containing the complete car object will be measured to ensure that large object that has been detected is not a passing pedestrian. Frame that pass from two stage of examination will be stored into the image file (*.jpg) and will be used for neural network training process after going through feature extraction process using gabor filter or instantly recognizable when the module that used is Automobile Recognition module. Examples of frames that are successful and not successful passing the segmentation phase are shown in Figure 4 and 5.

Figure 4. Example of frame that is successfully passing the segmentation phase

Figure 5. Example of frame that is failed to pass the segmentation phase

4.3 Feature Extraction Phase

In this phase, the result image from segmentation process will be convolved using gabor kernel for extracting the feature of image. The number of images generated from convolution process using gabor filter is equal to the number of gabor kernels that used to process an image. Gabor kernels are created based on orientation and frequency parameter that used. Suppose by using two types of orientation and five kinds of frequency will result ten kinds of gabor kernel. Image produced from this convolution process is called gabor response image. Example of this convolution process can be seen in Figure 6.

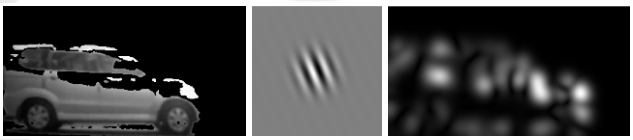


Figure 6. Image before gabor filtering, the gabor kernel and result image after the convolution process.

In this experiment, we used five different parameters of frequency ($n = 0, ..., 4$), namely: $W_n = 2^{-1}\pi, 2^{-1.5}\pi, 2^{-2}\pi, 2^{-2.5}\pi, 2^{-3}\pi$ and eight different parameters of orientation ($m = 0, ...,7$), namely: $\theta_m = 0^0$, 22.5^0, 45^0, 67.5^0, 90^0, 112.5^0, 135^0, 157.5^0. This is adjusted by the research conducted by Moreno et.al. [6]. The variations of gabor kernel are shown in Figure 7.

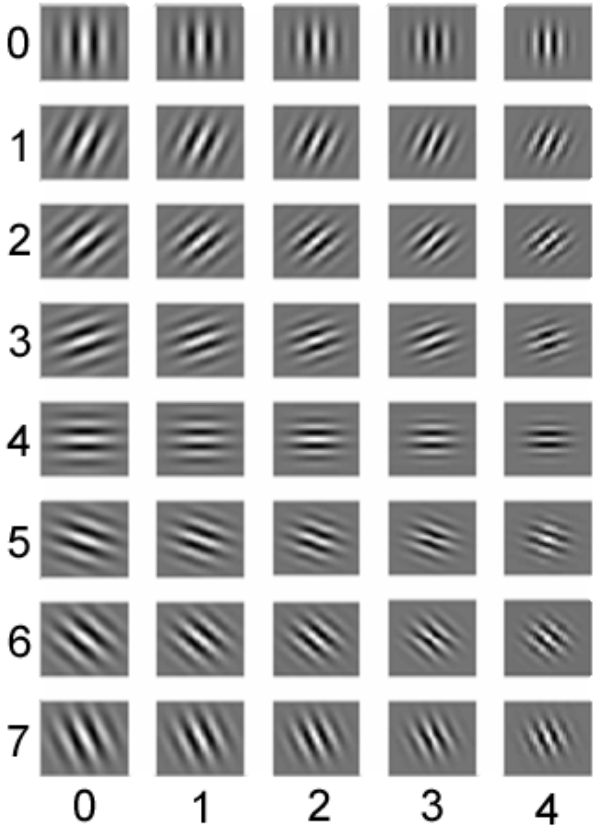


Figure 7. Forty types of gabor kernel [6].

4.4 Feed Forward Neural Network Design

Feed forward neural network architecture that is used is dynamic. Number of input neuron in input layer determined from the number of gabor response in feature extraction phase. After feature extraction process, this system will collect a sample point from result image that has been convolved by generated gabor kernel by using a grid. For example we used grid in the amount of ten, and from feature extraction phase we used four kinds of gabor kernel then the number of neurons in input layer is 10 x 10 x 4 = 400 neurons.

Hidden layer that is used is also dynamic, where we can change it freely according to the number of neuron in hidden layer and the number of hidden layer will be built.

Output layer consists of three neurons. All of three neurons represent the amount of automobile type that is processed. It is assumed the number of automobile-type in this world is less than or equals to seven kinds. Example formats of desire target: 000 = Sedan, 001 = City Car, 010 = MPV, etc.

4.5 Application Output

Output from this application is a list consists of automobile type that recognized successfully and their number. In real time mode, the number of automobile type is the number of vehicles that have been passed until then, since the application started. Meanwhile, if the mode is video file, this application will calculate the number and type of automobile that successfully identified. There are two kinds of displaying output: General List will display the recognition results based on the processing time, and Specific List will display the recognition results based on the automobile type. General and Specific List are shown in Figure 8.

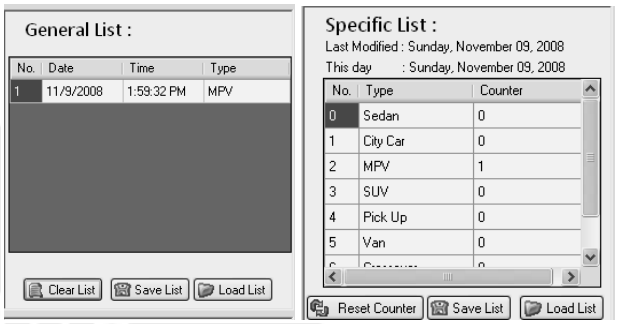


Figure 8. General List and Specific List examples.

5. TESTING AND ANALYSIS

The device specification used for testing is:

Processor : Pentium Core2Duo 2.4GHz

RAM : 2Gb DDR2

Harddisk : 250Gb

O/S : Windows XP SP2

Compiler : Ms. Visual Studio 2005 C++.Net

There are several kinds of experiments are performed, namely:

1. Test on the influence of the number of hidden layers and neurons in a hidden layer to the speed of training process until achieve convergence. The results of the tests are shown in Figure 9 and 10.

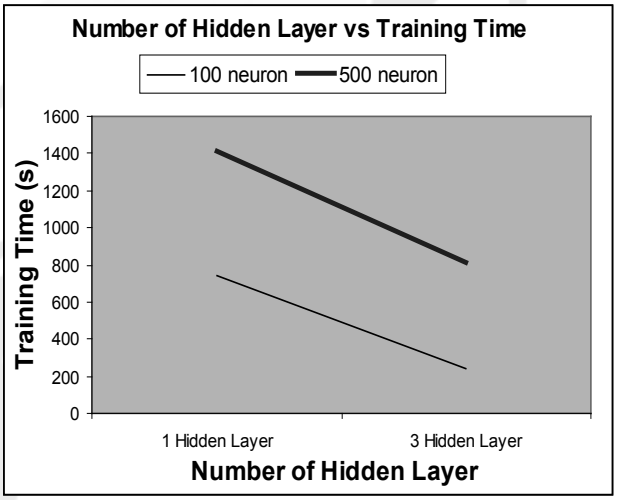


Figure 9. Number of hidden layer vs training time.

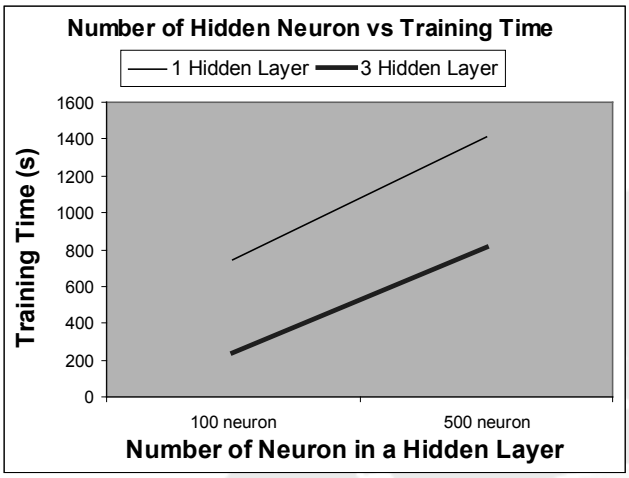


Figure 10. Number of hidden layer neurons vs training time.

From the experiments, we can conclude that the addition of hidden layers will speed up training time, while increasing number of neurons in the hidden layer will slow down the training time.

2. Tests for the error rate in recognition using data that has been trained to Neural Network. In this test, error rate that generated is 0% on the other words all the object has been recognized successfully. Gabor filter used frequency 0,1 and orientation 2,4.
3. Tests for the error rate in recognition using data that has not been trained to Neural Network. Gabor filter used frequency 0,1 and orientation 2,4. The results of the test are shown in Figure 11.

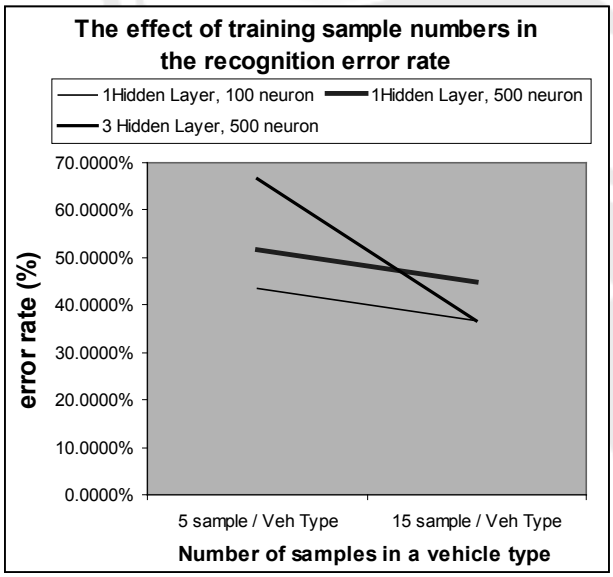


Figure 11. The effect of training sample numbers in the recognition error rate

From the tests, we can see that the quantity of training data for each automobile type can affect the accuracy of recognition for testing data that has not been trained yet. The percentage of lowest error rate for this test is 36%. From further analysis, we found that there is a type of automobile that is often wrongly recognized, it is the Crossover, because the shape of Crossover is similar to the shapes of Sedan, MPV or City Car. This application often made a mistake in recognized this type of automobile. If this type (Crossover) is removed from training data and testing data, then error rate will reduced to 24%.

4. The next test is to find the proper parameter setting of gabor filter. Various parameters of combination and frequency were attempted on the neural network with fixed configuration, one hidden layer and one hundred neurons in hidden layer. This back-propagation neural network setting is selected because it produces the lowest error rate (24%), with lowest number of hidden layer and neurons in hidden layer. From the test results, we found some parameter combinations of gabor filter which have the smallest recognition error rate, with percentage of error rate is 16% are:

   - Configuration 1: Orientation: 2; Frequency: 0, 2, 4, 6, 8
   - Configuration 2: Orientation: 3; Frequency: 0, 2, 4
   - Configuration 3: Orientation: 2, 3; Frequency: 0, 2, 4

5. The fifth test was used to see if the color of the car affects the rate of recognition. The result of the test is shown in Table 1.

Table 1. The result of automobile color test

Movie ID	Num. of Samples	Vehicle Color	*Error rate*
1	12	Gray(2), Silver(3), Black(3), Brown(2), Red, Blue	41.6667%
2	12	Blue(5), Silver(3), Black(2), Gray(2)	41.6667%
3	12	Silver(4), Blue(3), Green, Black, Brown, Gray, White	41.6667%

The error rates from several car colors that have been tested are same, which is 41.66%. It can be concluded that the color of the car does not affect the recognition rate.

6. The sixth test used to see if the lighting condition on the object that captured by a web cam can affect recognition rate or not. From the results of the tests known that if the lighting is too dark or too bright will cause the wrong silhouette object that will be processed as an input for the application. Examples of 'Too Dark' and 'Too Bright' object silhouette are shown in Figure 12 and 13.

Figure 12. An object and its 'too dark' object silhouette

Figure 13. An object and its 'too bright' object silhouette

6. CONCLUSION

In general, an application that is made in this study has achieved its purpose. By using gabor filters with orientation: 3 and the frequency: 0, 2, 4, and configuration of back-propagation neural network with one hidden layer and one hundred neurons in a hidden layer, we get the lowest error rate: 16%. This application is vulnerable to the effects of lighting when the lighting is too bright or too dark. This application is also experiencing difficulties in recognizing the shape of the hybrid-type automobile, such as crossover that is a mix of sedan and MPV or City Car.

7. ACKNOWLEDGMENTS

Our thanks to Mr. Resmana Lim and Mr. Thiang, who has helped us to get a better understanding of the research that they have done and also provide any important advice for the research we've done.

8. REFERENCES

[1] Burdick, H. E. 1997. Digital imaging: Theory and Applications. London: McGraw-Hill.

[2] Castleman, Kenneth R. 1996. Digital image processing. New Jersey: Prentice-Hall.

[3] Fadlisyah. 2007. Computer Vision dan Pengolahan Citra. Yogyakarta: Andi.

[4] Gonzales, Rafael C., and Woods, Richard E. 2002. Digital image processing 2nd edn. New Jersey: Prentice Hall.

[5] Haykin, S. 1999. Neural Networks: A Comprehensive Foundation, 2nd edn. Prentice Hall, Englewood Cliffs, NJ.

[6] Moreno, Plinio., Bernardino, Alexandre., and Santos-Victor, Jose. 2005. Gabor Parameter Selection for Local Feature Detection. IBPRIA - 2nd Iberian Conference on Pattern Recognition and Image Analysis, Estoril, Portugal, June 7-9, 2005.

[7] Negnevitsky, Michael. 2005. Artificial intelligence: a guide to intelligent systems 2nd edn. Pearson Education Limited. Addison-Wesley.

[8] Pratt, William K. 1991. Digital image processing 2nd edn. New York: John Wiley & Sons, Inc.

[9] Thiang, Guntoro, Andre Teguh., and Lim, Resmana. 2001. Type of Vehicle Recognition using Gabor Filter Representation and Template Matching Method. Proceeding, Seminar of Intelligent Technology and Its Applications (SITIA 2001).

[10] Tou, Jing Yi., Tay, Yong Haur., and Lau, Phooi Yee. 2007. Gabor Filters and Grey Level Co-occurrence Matrices in Texture Classification. Malaysia: Computer Vision & Intelligent Systems (CVIS) Group, Faculty of Information & Communication Technology Universiti Tunku Abdul Rahman (UTAR).

[11] Yap, W. H., Khalid, M., Yusof, R. 2007. Face Verification With Gabor Representation And Support Vector Machines. IEEE Proc. of the First Asia International Conference on Modeling & Simulation.

A Linear Graph and Genetic Algorithm Approach for Evolving Manipulator Modelling

Kok Kiong Tan
National University of Singapore
65-65162110
eletankk@nus.edu.sg

ABSTRACT
This paper addresses a methodology for modeling and designing evolutionary algorithms for 3 degrees of freedom (DOF) manipulators. The approach is based on linear graph modeling and genetic algorithm, where the latter is used to evaluate the fit of the linear graph model. The best fitness function of the genetic algorithm will be used to improve the performance of the manipulator.

Keywords
Linear graphs, Genetic algorithm, manipulator, modeling.

1. INTRODUCTION
Health monitoring system is an important tool for predicting, detecting, correcting and improving the system design to overcome malfunction problem. This is due to poor design could lead to poor performance of the system. The goal of this paper is to combine the health monitoring system and the evolutionary algorithm such that it could improve the system's performance, particularly the manipulator.

The manipulator is a combination of mechanical, electrical and control system. A conventional way of modeling the manipulator is by sequentially modeling each domain. Sequential domain modeling might not be optimal due to the dynamic interaction and energy transfer from each domain. Therefore, the algorithm of modeling which takes into consideration topological system structures and parameter values which could affect multi domain interaction and energy transfer could give a more optimal solution.

Seo et.al[2] combined bond graph modeling and genetic programming to evaluate the size and topology of mechatronics system. Bond graph modeling and genetic programming have been used for automated identification of a few mechatronic systems and a methodology has been developedfor multi domain design evolution by integrating health monitoring system, bond graph, genetic programming and the expert system.

Evolution of the system by combining some modeling methods and evolutionary algorithms has proven to be a potential tool to design an optimum solution of the system. However, this evolutionary method still requires determination of the component that can be modified such that it can achieve the desired performance. This component can be determined by using the expert system.

This paper presents an alternative way to find the optimum solution of the system by combining linear graph modeling and the genetic algorithm. The optimum solution can be used by a health monitoring system to improve its performance. A 3-DOF manipulator is employed as a plant to show the possibilities of the approach. Some optimum parameter values of the manipulator can be extracted and evaluated by using this method.

2. DESIGN FRAMEWORK
This paper takes advantage **of t**he flexibility in the linear graph modeling and the powerful evolutionary abilities of genetic algorithm to achieve the optimum model of the system.

Linear Graph modeling is a graphical representation of a system model which include the interconnection among its elements. The advantages of linear graph are [2]:

- Visualization of the system structure
- Identification of similarities between different types of the system domain
- Applicable for multi domain system
- Gives a unified approach to model multi functional devices

Since manipulator modeling needs a combination of mechanical, electrical and control system, linear graph modeling may be suitable for solving the modeling problem. Similar to the bond graph model, linear graph model describes the dynamic behavior of a system based on the principle of power and energy.

To get the equation from the linear graph, there are three types of equation needed to obtain an analytical model of the linear graph. They are:

- Constitutive equations for all elements that are not sources.
- Compatibility equations for the entire independent closed path.
- Continuity equations for all the independent junction of two or more branches.

On top of that, the linear graph must satisfy:

$$\ell = b - n + 1 \quad (1)$$

Where:

$\ell =$ *Number of loop*

$b =$ *Number of branch*

$n =$ *Number of nodes*

Genetic Algorithm (GA) is an evolutionary technique to find the optimum solution. GA requires genetic representation of the solution and fitness function of the system. The steps of GA are:

- Choose initial population. The initial population consists of some individuals. Each individual will represent the component of the system.
- Evaluate the fitness function of each population. The fitness function will be determined by the parameter performance of the system. The better the fitness function, the closer it is to the desired performance.
- Repeat this generation until termination. The generation will usually be terminated when the fitness function solution satisfies the minimum criteria of the design, maximum number of generation is reached, computation time is reached, etc.
- When the system is not terminated, GA will select the best fit individual for reproduction. Do some crossover and mutation to the best individual to give birth to offsprings. Finally, the new individuals will replace the least-fit population.

3. SYSTEM MODELLING

3-DOF manipulator is used as the plant of the engineering system. The manipulator consists of three revolute joints. The system can be seen in Figure 1.

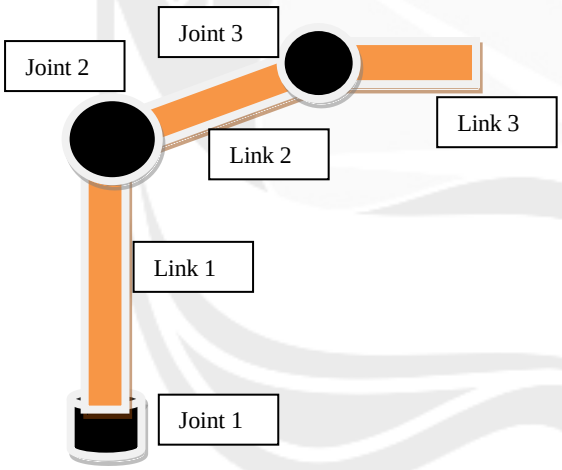


Figure 1. 3-DOF manipulator

The objective of this manipulator is to achieve the desired position from an arbitrary position. Each of the joint in the manipulator consists of an amplifier, a DC motor, gear transmission and a link to another joint. The movement of each joint will affect the performance of other joints. The manipulator is not under control by a control system. The objective is to design the manipulator so that it can achieve the desired position by changing the parameters. The linear graph of the manipulator can be seen in Figure 2.

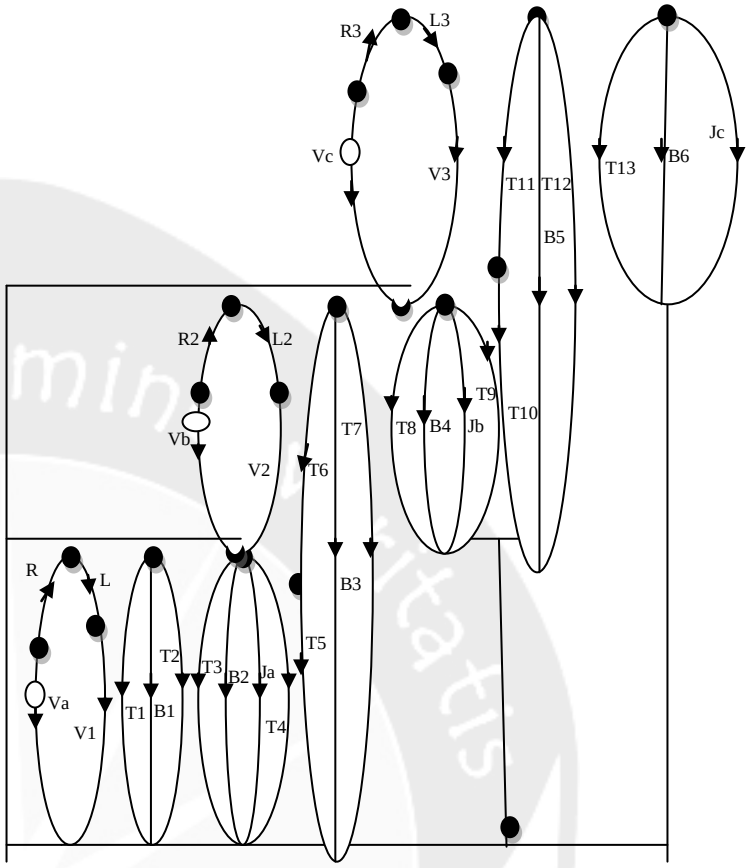


Figure 2. Linear graph modeling

As can be seen from Figure 2, number of loop equation = 17, number of branch = 34 , and number of nodes = 18 . From the topological equation in equation (1), the system satisfies the condition.

Moreover, it is possible to make a state space modeling of the system. The state variable can be chosen as ($i_1, i_2, i_3, \omega_a, \omega_b \omega_c$). The output from the motor will be positioned for link 1, link 2 and link 3($\omega_a, \omega_b \omega_c$). The input will be Voltage in link 1, link 2, and link 3. Therefore, the system could be described as MIMO (Multi Input Multi Output) systems which have 3 inputs and 3 outputs. With the assumption of no disturbance from the external system, the state space model is shown below:

$$B = \begin{bmatrix} \frac{1}{L} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & \frac{1}{L2} & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \frac{1}{L3} \\ 0 & 0 & 0 \end{bmatrix}$$

$$C = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

$$D = 0$$

$$x = \begin{bmatrix} \omega_a \\ i_1 \\ \omega_b \\ i_2 \\ \omega_c \\ i_3 \end{bmatrix} \qquad u = \begin{bmatrix} Va \\ Vb \\ Vc \end{bmatrix}$$

State Space Equation:

$$\dot{x} = Ax + Bu$$

$$y = Cx + D$$

4. SIMULATION RESULTS

There are 21 components used to design the system. The component can be shown in Table 1 below

Table 1. Component of the manipulator

Component	Value	Component	Value
R	1	R2	1
L	1	L2	1
N1	1	N2	1
Ka1	1	Ka2	1
Ja	1	Jb	1
B1	1	B3	1
B2	1	B4	1
R3	1	ra	1
L3	1	Jc	1
N3	1	B5	1
Ka3	1	B6	1
		rb	1

By substituting random numbers to each component, the performance of the system can be seen in Figure 3.

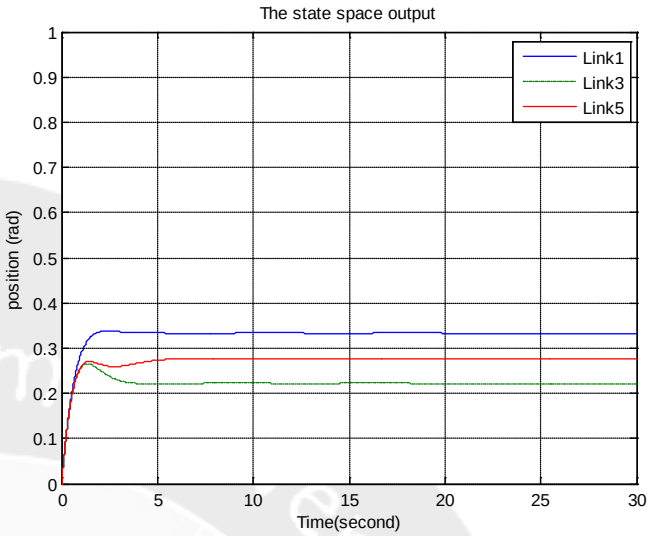


Figure 3. The performance before evolution

As can be seen in Figure 3, the performance of the system can be further improved. One possibility of improvement is by using evolutionary algorithms such as the Genetic Algorithm (GA). Each of the components from Table 1 will be chosen as the individuals in the GA. The variation value of every individual will be set as random 16 numbers which will vary from the minimum range value to the maximum range value. The variation value can be seen in Table 2.

Table 2. Variation value of each component

Component	Variation	Component	Variation
R	0-2	L2	0-1
L	0-1	N2	0-2
N1	0-2	Ka2	0-2
Ka1	0-2	Jb	0-2
Ja	0-2	B3	1
B1	1	B4	1
B2	1	ra	1
R3	0-2	Jc	0-2
L3	0-1	B5	1
N3	0-2	B6	1
Ka3	0-2	rb	1
R2	0-2		

The population and generation will be set as 100, and 500. The choice of the generation and population number may determine the performance of the evolving algorithm. Higher number of generation and population could improve the performance of the evolving algorithm. On the other hand, computation time will be longer due to complexity of the computation data.

Fitness function is also playing an important rule for GA. GA will keep the best value of the fitness function and replace the least-fit fitness function with another population. Another population can be generated by doing some crossovers and mutation methods. Crossover is a new population which came from some combination data from the best-fit population. The crossover from this paper is described in Figure 4 below.

individual 1				Individual 2			
1	1	1	1	0	0	0	0

crossover

individual 1	1	1	1	1
individual 2	0	0	0	0

result

new individual 1	1	0	1	0
new individual 2	0	1	0	1

Figure 4. Crossover

The fitness function formula is a specific solution for different cases. Therefore, the fitness function for each problem may be different; it is dependent on the parameter chosen. In this paper, the fitness function will be chosen as the summation of the squared error at every joint. The fitness function in this problem is chosen as:

$$fitness\ function = e_{ss1}^{\ 2} + e_{ss2}^{\ 2} + e_{ss3}^{\ 2} \quad (3)$$

Where:

$e_{ss1} = desired\ position - actual\ position\ of\ joint\ 1$

$e_{ss2} = desired\ position - actual\ position\ of\ joint\ 2$

$e_{ss3} = desired\ position - actual\ position\ of\ joint\ 3$

The best value of fitness function is defined as the minimum fitness function of the system. It is chosen due to the objectives of the design of manipulator, which is to make the error steady state of the manipulator to be smaller for each generation.

The simulation results when the population and generation are set to 100 can be seen in Figure 5.

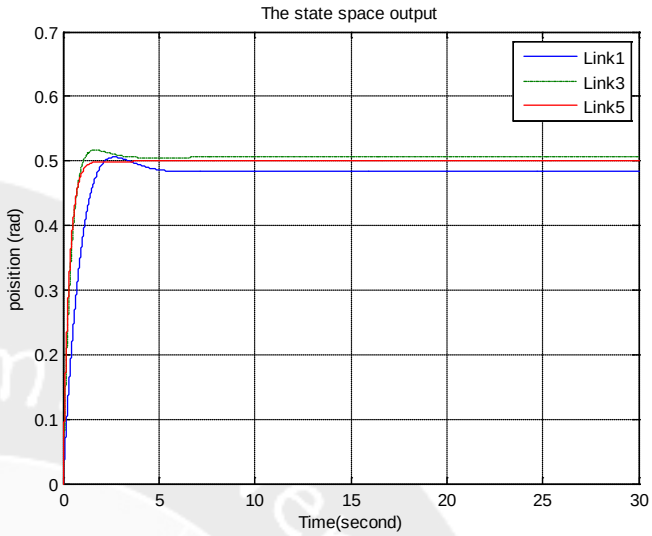


Figure 5. The position with 100 population and generation evaluation

As can be seen from the Figure 5, the simulation results shows that genetic algorithm works well. The design performance is getting closer to the desired performance. The fitness function evaluation is shown in Figure 6 below:

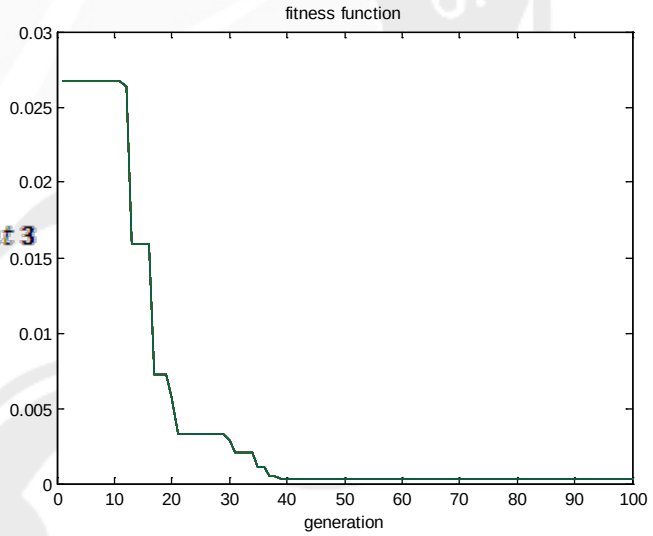


Figure 6. Fitness function after 100 generation evaluation

As can be seen form Figure 6, the fitness function is always being updated for the minimum value. However, the performance can still be improved by increasing the number of population and generation in the system. Therefore, another experiment using 500 population and generation is set. The results can be seen in Figure 7

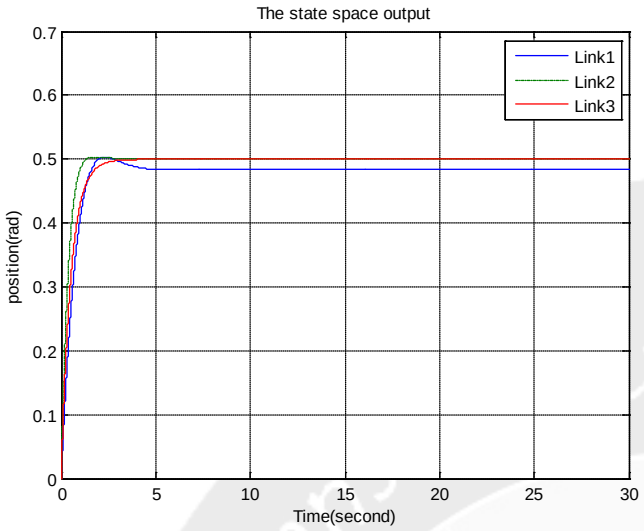


Figure 7. The position with 500 population and generation evaluation

As can be seen from Figure 7, the simulation result shows that the steady state error approaches zero. This is the case because the genetic algorithm continuously seek the minimum fitness function for each generation. The fitness function evaluation is shown in Figure 8.

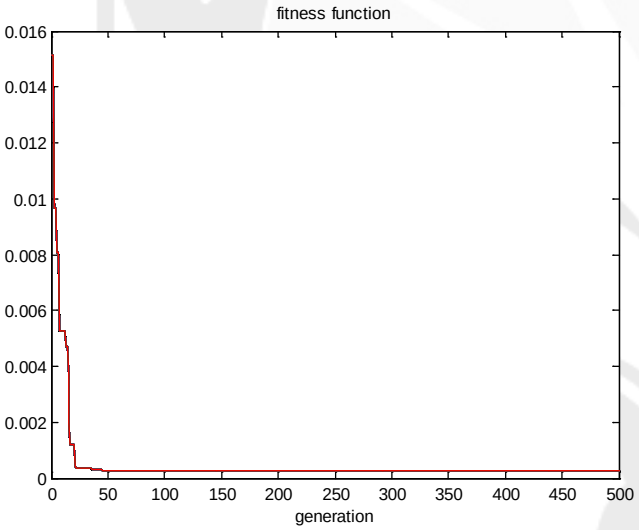


Figure 8. Fitness function after 500 generation evaluation

As can be seen, the genetic algorithm continuously seek the minimum value of the fitness function. A smaller fitness function implies a better performance and by minimizing this function, the actual performance will approach the desired performance. The updating parameter is shown in Table 3.

Table 3. Evaluating Parameter

Component	Value	Component	Value
R	0.5333	L2	0.4
L	0.8	N2	0.001
N1	0.001	Ka2	1.6
Ka1	1.4	Jb	0.6667
Ja	1.2	B3	1
B1	1	B4	1
B2	1	ra	1
R3	0.9333	Jc	2
L3	0.5333	B5	1
N3	0.001	B6	1
Ka3	1.6	rb	1
R2	0.8		

The updated parameters could be interpreted as a new value of the manipulator component. Simulation shows that, the new value could improve the performance of the system. However, the updated value may not be a fixed value which arises due to the possibility of multiple local optimal solutions for the problem.

4. CONCLUSION

Manipulator modeling which is concerned with systems containing multiple domain subsystems is a challenging problem due to the complexity with integration of each sub- system. A combination of linear graph and genetic algorithm is proposed in the paper to address this problem. The simulation result shows that the proposed method potentially can yield a good solution for the problem. Updating design could achieve the desired performance

5. REFERENCES

[1] Gamage, L.B. " A System Framework with On-line Monitoring and Evaluation for Design Evolution of Engineering Systems"

[2] K. Seo, Z. Fan, J. Hu, E.D. Goodman, and R.C. Rosenberg, "Toward a unified and automated design methodology for multi domain dynamic systems using bond graphs and genetic programming,"*Mechatronics*, Vol. 13, No. 8-9, pp. 851-885, 2003

[3] Said, H.Y. "On Genetic Algorithm and Their Application" Handbook of Statistic, Vol.24. 2005

[4] Riechman, T. "Genetic Algorithm Learning and Evolutionary Games". Journal of Economic Dynamic & Control 25. 2001

[5] Skinner M. "Genetic Algorithm Overview". http://geneticalgorithms.ai-depot.com/Tutorial/Overview.html

Comparing Genetic and Ant System Algorithm in Course Timetabling Problem

Djasli Djamarus
Informatics Department
Trisakti University
Jakarta 11440, Indonesia
+62 21 5631003
djasli@trisakti.ac.id, djamarus@gmail.com

ABSTRACT
This paper models the Course Timetabling Problem as a set of tuples each of which consists of four entities, i.e. lecturers, courses, rooms and time-slots that have to be matched in order to construct the intended course timetable. This model considers lecturers' time preferences and their expertise as the constraint of the problem. A bipartite graph that connects the four entities is used as the path of ant movement. In this experiment two meta-heuristics algorithm, Genetic and Ant System Algorithm is applied to the problem, and the results are compared.

Keywords
Ant System Algorithm, Course Timetabling Problem, Genetic Algorithm, Meta-heuristic.

1. INTRODUCTION
Genetic Algorithm and Ant System Algorithm are two among meta-heuristic algorithms that have been used to solve some NP (Non deterministic Polynomial) problems, such as TSP [6][12][20].

In educational institutions, the NP problem appears in activity to construct educational timetables, such as school timetable, examination timetable and course timetable [18]. School timetable is used in intermediate level educational institutions while the other two are required in the higher level educational institutions, especially the ones that apply the credit unit system.

Course timetabling, as the name imply, is an activity to produce course timetable that can be used operationally by an educational institution for a certain period of time usually 4 up to 6 months, known as semester.

Although this activity repeatedly happens in every semester, not much educational institution constructs its course timetable in a fully automatic manner. This reality condition happened not only due to so many variations in the way of constructing course timetable, but also naturally this problem, except for a limited number of data, is required a very long time to be solved by deterministic algorithm. The first implies that so far there is no single application program can be used by all educational institution to construct the course timetable, mean while the second suggests the schedulers to find alternative algorithms in solving their problem, such as stochastic approach, in order to come with satisfying solution in a reasonable running time. In fact the course timetabling problem has been studied extensively and it has been noted that no less than 700 academic writings have been published in this area [3] since it was first introduced by Gotlieb in 1962 [8]. Based on literature survey, it is known that The Genetic Algorithm has been used by [2][14] and [1], while the Ant System Algorithm is used by [19] and [13].

The rest of this paper is organized as follows. Section 2 describes the course timetabling problem, while Section 3 and 4 describes the two meta-heuristic algorithms used in this research. Problem Model is described in Section 5, follow by Experimentation and results in Section 6 and finally concluding remarks are given in the last section.

2. COURSE TIMETABLING PROBLEM
The course timetabling problem is defined as a sub-class of assignment problem for which the events take place at educational institutions [22]. The goal of the problem is to arrange events, each of which conducted by lecturer in a certain room in a definite period of time to be offered for students [21].

Difficulties of the course timetabling problem usually come from its constraint that must be satisfied. There are two types of constraint that must be obeyed by the scheduler. The first is called hard constraint that is mandatory to be followed otherwise the schedule is not workable. The existence of a unique resource, such as a person, in a certain time usually has to be considered as a hard constraint. The second is soft constraint that will be better if followed by the scheduler. Preference for something related to arrangement of resources could be considered as a soft constraint.

Basically there are two approaches that could be followed by the timetable officer in order to construct the timetable [10]. The first predicts what courses should be offered to students in order to satisfy the student requirement, and then assigned some resources to the course sections. The second allows students to select any course sections, and then arrange the sections so that everyone happy with the timetable.

This research follow the first approach so that the course timetable is represented as a set of tuples each of which consists of course, lecturer, contiguous time-slots and room that imply a relationship as in a sentence "the course will be conducted by the lecturer every week in between the time-slots in the stated room."

Theoretically, an exact or perfect solution for such timetabling problem can be found only by evaluating every possible solution candidate systematically. However this exhaustive deterministic method can be applied for a limited number of data only. The real course timetabling problem usually has much data to process, so that this method will require a very long time to wait for the program to come up with the perfect result [16]. The relation between amount of data and the computer running time can be

expressed as a combinatorial function that grow much faster than polynomial function.

As the NP Problem, currently the most promising approaches to solve the course timetabling problem are local search and constraint programming approach. Usually these methods are included in a class of algorithm called meta-heuristic [4][5][9]. The local search starts by proposing an arbitrary solution, and then looks for better solution from the neighborhood repeatedly, among this approach are taboo search, simulated annealing, and genetic algorithm. The latter declares the problem as a set of constraints that define relations among variables that must be obeyed in search of a solution [17]. This approach usually starts by defining the solution as an empty set and then gradually includes some components that bear with the defined constraints into the solution [4] also called this approach as the constructive approach. The recent meta-heuristic algorithm that uses the constructive approach in solving the problem is the Ant System Algorithm.

```
01 : Construct Random Population
02 : Select the best Individual as Solution
03 : Calculate Fitness Value
04 : Perform Reproduction peration
05 : Perform Cross over operation
06 : Perform Mutation operation
07 : Select the new best Individual
08 : If new best Individual better than Solution replace
     Solution with new best Individual
09 : If stopping condition is not met Go to step 03
```

Figure 1. Pseudocode of the genetic algorithm.

3. GENETIC ALGORITHM

Genetic Algorithm is kind of meta-heuristic algorithm that inspired by the mechanic of biological selection introduced by Charles Darwin, in order to search the fittest individual within a population. An individual is considered as a collection of chromosomes each of which constituted by genes. The characteristic of individuals that depends on composition of chromosomes is known as the fitness value of the individuals that reflects the quality of individuals. The evolution of individual is controlled by three basic processes of the Genetic Algorithm, namely reproduction, cross over and mutation [15]. The Genetic Algorithm mimics all of these processes step by step as in pseudocode written in Figure 1.

For the course timetable problem, the population is some sets of possible course timetable, while the chromosome is a course timetable element. The Genetic Algorithm is used to search the fittest individual that represents the best possible course timetable. The chosen timetable should have the least contradiction to the constraints. It means that the fitness function will depend on the violation of individuals to the constraints. In the iteration process, one of the termination conditions will be the fitness value of any individual that is close enough to the satisfying fitness value.

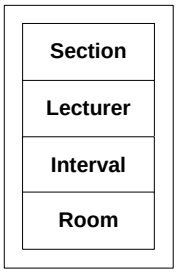


Figure 2. Chromosome.

The reproduction process can be done straight forward based on the fitness value of each individual in the population. However the cross over and mutation process may cause one chromosome has its duplicate in the same individual, which means to worsen the individual. This unexpected result must be catered by other process to improve the individual.

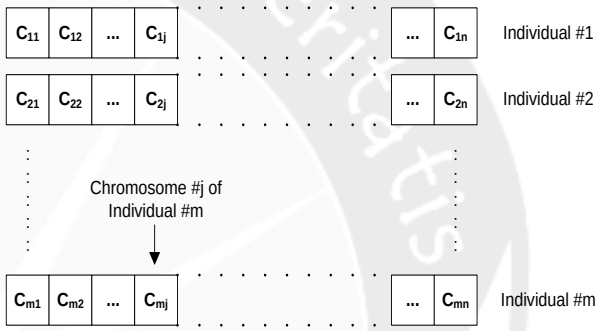


Figure 3. Population of genetic algorithm.

In order to implement the Genetic Algorithm into a program, the chromosome is represented in a data structure as in Figure 2, while the data structure of the population is shown as in Figure 3.

4. ANT SYSTEM ALGORITHM

The Ant System Algorithm is another meta-heuristic algorithm that also inspired by biological phenomenon. This algorithm firstly introduce by [7]. The system was inspired by the fact that ants as a colony manage to find the shortest path to reach their food. Each time an ant moves from one place to another place it leaves some substances that called pheromone on its trail as the communication medium. This chemical substance functions as positive feedback for the colony, so that the more ants passing a particular path the more pheromone left on that path and the more probability the colony to follow the path.

If there are some paths between nest and a place of food and there is a colony of ants move back and forth from the nest to the place of food, in the beginning the number of ants passing the path will distribute more or less evenly.

It can be understood that in a certain period of time ants passing the shortest path will be more frequent in comparison with other paths. It means that the colony will put more pheromone in the shortest path. Logically from time to time the shortest path will be passed by more ants and eventually the shortest path will become the only path used by the ant colony. Figure 4 illustrates the ant colony in discovering the shortest path with only two alternative paths to follow.

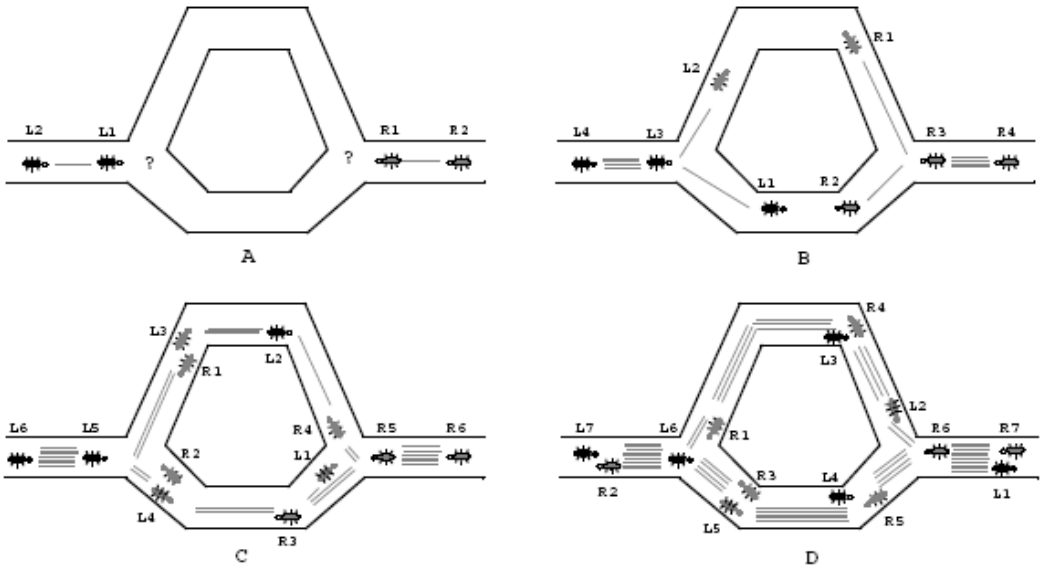


Figure 4. Colony of ants finding the shortest path [12].

As an algorithm that used construction approach finding solution, the Ant System Algorithm collects one by one element to construct the timetable. In the course timetabling problem, the algorithm begins the timetable construction with an empty timetable. For every iteration step, the algorithm will try to collect a timetable element that does not cause any conflicts with any other elements that have been selected previously.

For that reason, to construct the propose timetable the algorithm requires a comprehensive checking activity in order to prevent constraint violation when adding the selected element into the proposed timetable. With this effort all elements in the proposed timetable are conflict free to each other at any step. Pseudocode of the Ant System Algorithm can be written as in Figure 5.

```
01:  Initialize Parameters and Pheromone Trail
02:  Construct Ant Solution
03:  Update Pheromone Trail
04:  Perform centralize Action
05:  If stopping condition is not met Go to step 02
```

Figure 5. Pseudocode of the ant system algorithm.

5. THE PROBLEM MODEL

The course timetabling problem in this problem consists of four entities namely course, C = {c_1, c_2, … c_i}, lecturer, L = {l_1, l_2, … l_j}, time slot, T = {t_1, t_2, … t_m} and room, R = {r_1, r_2, … r_n}. Each course timetable requires one element of course, ci, one element of lecturer, l_j, one or more element of time-slots, t_k ... t_m where k < m, and one element of room, r_n.

The maximum sections of a certain course and the maximum teaching hours of the lecturer are taken as the hard constraint, while the expertise level and the priority level of time-slots for the lecturer are considered as the soft constraint.

To employ the Genetic Algorithm for the course timetabling problem, the propose solution must be structured in the form of individual that consists of chromosomes. In this model, each chromosome represents a tuple <Course, Lecturer, Intervals, Room>, where each component is assumed as a gene that represents one of the entities.

The reproduction process can be done straight forward based on the fitness value of each individual in the population. However the cross over and mutation process may cause one chromosome has its duplicate in the same individual, which means to worsen the individual. This unexpected result must be catered by other process to improve the individual.

For the Ant System Algorithm, the problem is modeled as a bipartite graph G (V, E), where V is a set of vertices represent the entities course, lecturer, time-slot and room that involve in the problem. E is a set of edges which represents the possible relationship between those entities. Relations between course c_i and lecturer l_j indicates lecturer l_j is willing and capable to handle course c_i, relation between lecturer l_j and time-slot t_m indicates availability of lecturer l_j to conduct in time-slot t_m, and relation between time-slot t_m and room r_n indicates the availability of room r_n in the t_m time-slot. The graph will be used by the ant to move among vertices to search tuples required by the Genetic Algorithm. By adding one vertex at both ends, each of which as nest and food, the graph as in Figure 6 is very similar to the original Ant System model.

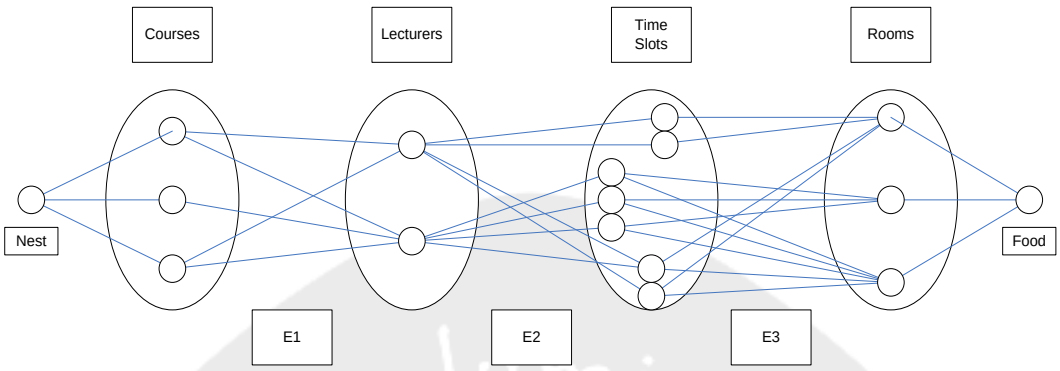


Figure 6. Graph model for ant system algorithm.

This computation model gives a freedom to choose a path that connect the nest to the food, through any vertex that construct a tuple <Course, Lecturer, Intervals, Room> as a candidate of timetable element. The element that represented by sequence of edges in the path between nest and food, will be selected based on the number of pheromone deposited by the ant that moves from nest to food. Therefore, before the path selection is performed, there should be a process directed by some heuristic factors to construct the pheromone trail [11].

Natural hard constraints that prevent a lecturer to do more than one job in a certain time-slots and the multiple used of a room in a certain time are represented by weight of the relation that is limited to 1. Hard constraints are handled by decreasing the weight of edges by 1 every time the edge is selected, so that there is no negative weight in the graph. The soft constraints that are preferable to be conformed are managed by an approach of the pheromone trail construction.

Table 1. Course entity

Course	Section	Credit Unit
C00203	11	2
C00258	14	2
C00329	9	3
C00330	10	3
C00355	13	3

6. EXPERIMENT AND RESULTS

The data used for this problem consist of Courses, Lecturers, Time-slots and Rooms entities as well as relationship among them in order to construct a workable timetable. In this case study 2 random data sets are used. The first consists of 50 courses, 100 lecturers, 60 time-slots (12 time-slots/day for 5 days/week) and 20 rooms, aim to construct 343 sections, while the other having 114 courses, 336 lecturers, 72 time-slots (12 time-slots/day for 6 days/week) and 80 rooms, in order to construct 1729 sections.

A sample of the input data format is presented in Tables 1 and 2. The number of section required to be opened for a certain course is indicated by 'Section' in Table 1. For example, course IIE203 requires 4 sections to be opened and each section would require 2 contiguous time-slots. The number of time-slots needed for each course is represented by 'Credit Unit'. The teaching load for each lecturer is indicated by the number of maximum sections the lecturer could fulfilled, while maximum total lecturing time for each lecturer is indicated by 'Duration of Lecture' as depicted in Table 2.

Table 2. Lecturer entity

Lecturer	Load (Section)	Max Hours
L1002	5	13
L1007	2	5
L1008	5	14
L1025	3	9
L1033	4	10

Data for time-slots are presented as T101, T102, …., , T511, T512. For example, T101 represents the first time-slot on Monday while T512 represents the twelfth time-slots on Friday. There are 5 working days in a week and 12 time-slots in a day. Rooms are numbered from 1 to 20 since there twenty available rooms in this case. A sample of the relationship between Courses and Lecturers is depicted in Table 3 while Table 4 presents a sample of the relationship between lecturers and time-slots.

Table 3. Lecturer entity

Course	Lecturer	Max Section	Expert Level
C00203	L1003	1	2
C00203	L1022	1	1
C00205	L1008	1	1
C00233	L1010	2	1
C00252	L1045	2	3
C00309	L1001	1	1
C00312	L1094	2	1
C00313	L1127	2	1
C00316	L1104	2	2
C00346	L1146	2	1

In the program that is developed based on the Genetic Algorithm, the number of individuals in the population is set to 20. Both

mutating and crossing over probability are set to 0.1. The program that based on Ant System Algorithm is set up so that the number of pheromone on the edges is limited in the range of 10 (minimum) up to 1000 (maximum). The pheromone on each edges of success path will be increased by a number of 10 * q1 / (1 + q2), where q1 is the capacity of the destination vertex, and q2 is the sum of out-degree of source vertex and in-degree of destination vertex. Evaporation rate is assumed constant as much as 0.1%.

Table 4. Lecturer entity

Lecturer	Time-slot	Unit	Priority
L1001	T101	3	1
L1001	T201	3	1
L1003	T101	3	1
L1003	T501	3	5
L1010	T404	2	4
L1045	T204	3	2
L1045	T503	3	5
L1094	T401	2	3
L1104	T106	3	2
L1127	T401	3	4

The results are averaged and compared to each other as well as to the result of Greedy Algorithm as in Figure 7. The vertical axis shows the average number of sections can be scheduled by each algorithm.

7. CONCLUSION

From the result, it can be seen that both Genetic and Ant System Algorithm show better result compare to a greedy algorithm in constructing course timetable using the proposed computation model. However, the improvement is getting less in a problem that having more data input.

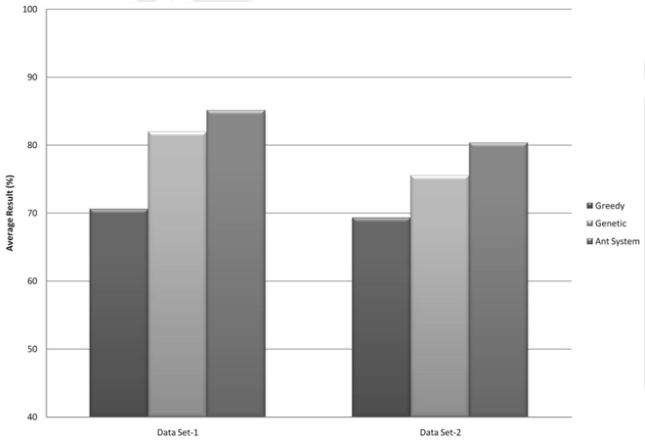


Figure 5. Result Comparisons of the Algorithms.

Another thing can be seen from the experiments is that the result of Ant System Algorithm to the Greedy shows more improvement compare to the Genetic Algorithm.

The experiment also shows that the more input data used as the input, in percentage, the less average sections can be scheduled, meaning that the more difficult to find the satisfying timetable.

8. REFERENCES

[1] Abdullah, S., and H. Turabieh. "Generating University Course Timetable Using Genetic Algorithms and Local Search." *Third International Conference on Convergence and Hybrid Information Technology*. Busan, Republic of Korea, 2008.

[2] Bambrick, Leon. *Lecture timetabling using genetic algorithms*. Brisbane: The University of Queensland, 1997.

[3] Bardadym, V A. "Computer-Aided School and University Timetabling: The New Wave." *First International Conference on the Practice and Theory of Automated Timetabling (ICPTAT '95)*. 1995.

[4] Blum, C., and S. Roli. "Metaheuristics in Combinatorial Optimization: Overview and Conceptual Comparison." *ACM Computing Surveys* 35, no. 3 (2003): 268 - 308.

[5] Burke, E., and S. Petrovic. "Recent Research Direction in Automated Timetabling." *European Journal of Operational Research – EJOR* 140, no. 2 (2002): 266 - 280.

[6] Chatterjee, S., C. Carrera, and L. A. Lynch. "Genetic Algorithms and Traveling Salesman Problems." *European Journal of Operational Research – EJOR* 93, no. 3 (1996): 490 - 510.

[7] Colorni, A., M. Dorigo, and V. Maniezzo. "Distributed Optimization by Ant Colony." *First European Conference on Artificial Life*. 1992.

[8] Csima, J., and C. C. Gotlieb. "Test on Computer Method for Constructing School Timetables." *Communications of the ACM* 7, no. 3 (1964): 160 – 163.

[9] Digalakis, J., and K. Margaritis. "A Parallel Memetic Algorithm for Solving Optimization Problems." *4th Metaheuristics International Conference (MIC'2001)*. Porto, Portugal, 2001.

[10] Djamarus, Djasli. *Enhancement of Ant System Algorithm for Course Timetabling Problem.* Kedah: Universiti Utara Malaysia, 2009.

[11] Djamarus, Djasli, and Ku Ruhana Ku-Mahamud. "Heuristic Factors in Ant System Algorithm for Course Timetabling Problem." *Ninth International Conference on Intelligent Systems Design and Applications*. Pisa, Italy: IEEE Computer Society, 2009. 232 - 236.

[12] Dorigo, M., and L. M. Gambardella. "Ant Colony System: A Cooperative Learning Approach to the Traveling Salesman Problem." *IEEE Transactions on Evolutionary Computation* 1, no. 1 (1997).

[13] Ejaz, N., and M. Y. Javed. "A Hybrid Approach for Course Scheduling Inspired by Die-hard Co-operative Ant Behavior." *IEEE International Conference on Automation and Logistics*. Jinan, China, 2007.

[14] Erben, W., and J. Keppler. "A Genetic Algorithm Solving a Weekly Course-Timetabling Problem." *Practice and Theory of Automated Timetabling First International Conference.* Edinburgh, UK, 1995.

[15] Goldberg, David E. *Genetic Algorithm in Serach, Optimization and Machine Learning.* Reading: Addison Wesley, 1989.

[16] Huang, X. "A Polynomial-Time Algorithm for Solving NP-Hard Problems in Practice." *ACM SIGACT News* 34, no. 1 (2003): 101 - 108.

[17] Marte, M. *Models and Algorithms for School Timetabling – A Constraint Programming Approach.* Munchen: Fakultat fur Mathematik, Informatik und Statistik, Ludwig-Maximilians Universitat, 2002.

[18] Schaerf, A. "A Survey of Automated Timetabling. Artificial Intelligence Review." 13, no. 2 (1999): 87 - 127.

[19] Socha, K., J. Knowles, and M. Sampels. "A Max-min Ant System for the University Course Timetabling Problem." *ANTS 2002 – Third International Workshop on Ant Algorithms.* 2002.

[20] Stutzle, T., and H. Hoos. "Max-min Ant System and Local Search for the Traveling Salesman Problem." *IEEE International Conference on Evolutionary Programming.* 1998.

[21] Tripathy, A. "School Timetabling – A Case in Large Binary Integer Linear Programming." *Management Science* 30, no. 12 (1984): 1473 - 1489.

[22] Willemen, R. *School Timetable Construction – Algorithms and Complexity.* Eindhoven University of Technology, 2002.

Gas Distribution Network Optimization with Genetic Algorithm

K. A. Sidarto
Department of Mathematics
Institut Teknologi Bandung
Jl. Ganesha 10, Bandung
+62 22 2508126
sidarto@math.itb.ac.id

L. S. Riza
Study program of Computer Science
Universitas Pendidikan Indonesia, Bandung
Jl. Setiabudhi 229, Bandung
+62 8157025880
lala_s_riza@yahoo.com

C. K. Widita
Research Consortium OPPINET
Institut Teknologi Bandung
Jl. Ganesha 10, Bandung
+62 22 2508126
chasanah.k.widita@gmail.com

F. Haryadi
Research Consortium OPPINET
Institut Teknologi Bandung
Jl. Ganesha 10, Bandung
+62 22 2508126
fb_haryadi@yahoo.com

ABSTRACT

Now days, natural gas plays an important role as a source of clean energy. The addition of gas consumption generally will require the new design and construction of gas pipelines. In this regard, the pipe diameter optimization process by considering the technical specifications is a must. Using the obtained optimum gas pipeline's diameter, the investment cost and gas operations can be minimized. Gas distribution pipeline network consists of nodes that represent points of consumers and suppliers connected by a pipe. Assuming the gas flow in a *steady state*, pipe networks are modeled into a nonlinear equation system from gas flow equations in the pipe. This model system solved by Genetic Algorithm to obtain the optimum gas pipeline's diameter with an investment cost of the pipe system as an objective function and specification of pressure on a node as a constraint. The optimization process is optimization of pipe specifications which available on the market (ANSI / ASME) with 64 kinds of diameter with range from 3 to 16 inch. At the end of the paper, a case study the optimization of gas pipe diameter in the region X is presented. From these case studies can be concluded that the Genetic Algorithm can determine the optimum pipe diameter which gives the lowest investment costs while still consistent to the technical specifications that have been determined.

Keywords

Genetic Algorithm, gas distribution pipelines, optimization of gas pipe diameter.

1. INTRODUCTION

Currently, natural gas plays an important role in providing clean energy for the community. With the increasing gas demand, network development required a new gas pipeline to meet the needs of consumers and to connect the dots of new customers. To perform the design and construction or expansion of gas distribution pipelines, pipe diameter optimization process must be done to minimize the investment cost. On the other hand, the gas company also has a responsibility to meet the needs of consumers with gas to the pressure and flow rates that have been agreed in the contract. Therefore, the optimization is an optimization performed with specific limitations for pressure and flow rate that has been agreed in the contract.

This study focused on determining optimum pipe diameter and pressure distribution which gives the pipe's minimum investment cost, and also perform economics calculations model (consisting of investment costs, coating costs, installation costs and operational costs of pipes). Optimization pipe diameter must also consider the balance of the pressure distribution on the pipeline. Gas distribution pipeline network is modeled as the pipes that connect some point the gas supplier to the consumer points assuming the gas flow in a steady state.

The system model which used to represent the gas distribution system is a method of balancing the gas flow in the pipe. Stoner [1] was the first time using this model for large networks. Stoner proposed steady state model written from the substituted gas correlation into the flow balance model, thus the nonlinear equation system is obtained. Then the nonlinear equation system will be solved by Genetic Algorithm [2], [3]. Diameter optimization with Genetic Algorithm based on the specifications of pipe which is available on the market and the allowed pressure distribution (ANSI / ASME [4]). After getting the optimum pipe diameter, the economic model i.e. investment costs, coating costs, installation costs, and operational pipe costs will be calculated.

2. METHODOLOGY

The problems solved step by step follows the flowchart below.

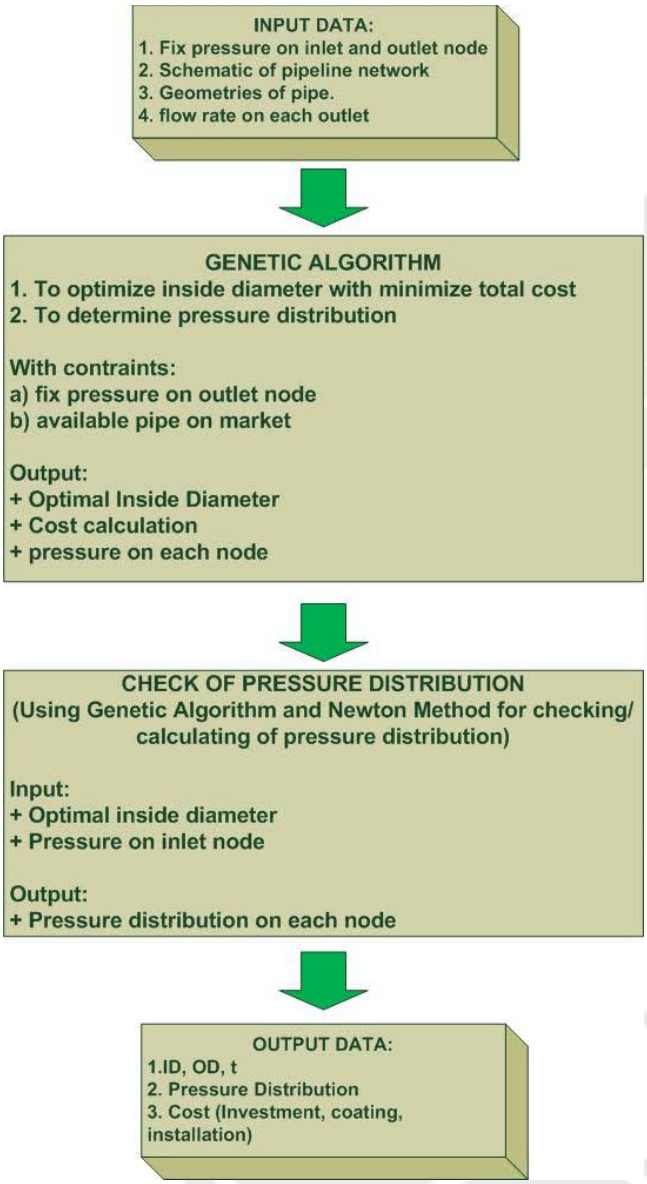


Figure 1. Step-by-step for solving the problem

As shown in the picture above, there are three main parts i.e. the Genetic Algorithm application to obtain the optimum pipe diameter, the calculation of economic models, and the application of Newton's method for calculating the pressure distribution. This paper is focused on explaining the application of Genetic Algorithm and cost calculation, while Newton's method for calculating the pressure distribution has been discussed in the previous paper [5].

2.1 Model Formulation

The system model used to represent the gas distribution system is the flow balance method, i.e. the volume of gas flowing into the system must be equal to the volume of gas that came out of the system. Gas flow correlation used in this study is the Panhandle A correlation.

A pipe connecting node *i* and node *j* has a length L_{ii} (*mile*) and inside diameter ID_{ii} (*inch*). Pipeline system is assumed in constant conditions or *steady-state* with the gas temperature *T*, specific gravity *G*, and pipe efficiency *E*. The flow from *i* to *j* is expressed as a positive flow. Gas flow rate has units of *MMSCFD* and gas pressure has units of *psia*. For horizontal flow, the Panhandle A correlation is given by [6], [7]:

$$Q_{ij} = S_{ij}\frac{C\,E\,ID_{ij}^{\,2.6128}\left(\left|P_i^2 - P_j^2\right|\right)^{0.5394}}{S\,Gg^{0.4606}\,T^{0.5394}\,L_{ij}^{0.5394}} \quad (1)$$

with Q_{ii} is the flow rate of gas in the pipe which connecting nodes *i* and *j*. P_i and P_i are the pressure at node *i* and node *j*, respectively. *C* is a constant correlation. To simplify the problem, it is assumed that all segments of the pipe work in conditions of $T = 60$, $G = 0.6$ and $E = 0.92$. Thus Panhandle A correlation can be simplified to:

$$Q_{ij} = \frac{K\,D_{ij}^{\,2,6182}\left(P_i^2 - P_j^{\,2}\right)^{0,5394}}{L_{ij}^{\,0,5394}} \quad (2)$$

with $K = 8.2634*10^{-4}$.

Flow balancing model is built by applying the analogy of Kirchhoff's law in electricity, so for a point *m*, the continuity equation obtained as follows[4]:

$$f_m = Q_{jm} - Q_{mk} + QN_m = 0 \quad (3)$$

the index of *Q* shows connectedness while the + / - indicates direction of flow. While f_m is a nonlinear equation at nodes *m* and represents the flow imbalance at some point, so it is zero if the system is in a state of balance. If the gas pipeline has a 10 point must be connected by pipeline segment then 10 nonlinear equations will be obtained.

As mentioned earlier, the optimization process carried out to obtain the minimum investment with prescribed constrain of pressure. Investment formula is as follows.

$$CIP_{ij} = \frac{10.68\left(OD_{ij} - t_{ij}\right)t_{ij}\,L_{ij}\,C_{pipe}\,5280}{2000} \quad (4)$$

With the cost of investment CIP_{ii} pipe, outside diameter OD_{ii}, t_{ii} is the wall thickness of the pipe between nodes *i* and *j*, and C_{pipe} is pipe cost per ton. Total cost of investment in the whole system of pipes is as follows.

$$CIP_{total} = \frac{10.68\left(\sum_{i,j;i\neq j}\left(OD_{ij} - t_{ij}\right)t_{ij}\,L_{ij}\right)C_{pipe}\,5280}{2000} \quad (5)$$

L_{ii} is the length of the pipe between nodes *i* and *j* which already known.

In addition, another economies models calculated by the following formula.

Total of coating cost:

$$C_{coat}\sum_{i,j;i\neq j} L_{ij} \quad (6)$$

C_{coat} is coating cost per mile.

Total of installation cost:

$$C_{inst} \sum_{i,j;i\neq j} L_{ij} D_{ij} \quad (7)$$

C_{inst} is installation cost per inch per mile.

Total of operational cost, assumed 4% of total of investment cost and total of coating cost.

$$0.04\left(\frac{10.68\left(\sum_{i,j;i\neq j}\left(OD_{ij} - t_{ij}\right)t_{ij}\,L_{ij}\right)C_{pipe}\,5280}{2000}\right) + C_{coat}\sum_{i,j;i\neq j} L_{ij} \quad (8)$$

So, to minimize the total of pipe investment cost in steady state condition, nonlinear optimization problem below has to be solved.

Minimize

$$CIP_{total} = \frac{10.68\left(\sum_{i,j;i\neq j}\left(OD_{ij} - t_{ij}\right)t_{ij}\,L_{ij}\right)C_{pipe}\,5280}{2000} \quad (9)$$

subject to

$$F(x) = \frac{1}{1+\|f(x)\|} = 0, \quad wh\ re\ \|f(x)\| = \sqrt{f_1^2(x) + f_2^2(x) + \ldots + f_n^2(x)} \quad (10)$$

2.2 Computation Methods

Genetic Algorithm (AG) is a random search algorithm based on natural selection and genetics mechanisms. AG developed by John Holland at the University of Michigan in 1975.

AG working in a set of candidate solutions called population. Each candidate solution is called an individual in the population. Usually, the individual is represented by a binary string. Any individual or string is mapped in a fitness value that represents the individual's level of performance. Each individual in the population will be subject to an operation to improve his fitness value. The operation is the selection or reproduction, crossover or cross-breeding, and mutation. Genetic Algorithm can be described as flowchart below.

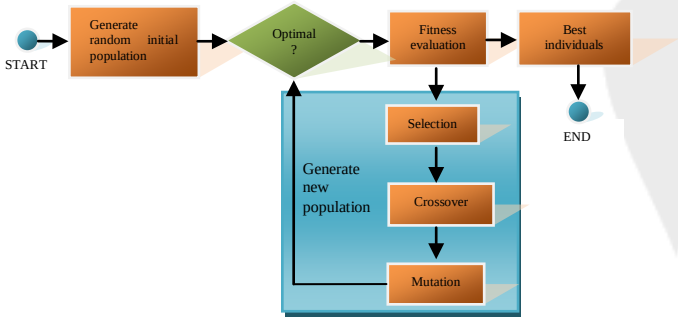


Figure 2. Flowchart of Genetic Algorithm

The basic steps in a Genetic Algorithm are [8]:

1. Generate randomly an initial population of chromosomes.
2. Calculate the fitness, defined according to some specified criteria, of all the members of the population and select individuals for the reproduction process. The fittest are given a greater probability of reproducing in proportion to the value of their fitness.
3. Apply the genetic operators of crossover and mutation to the selected individuals to create new individuals and thus a new generation. Crossover exchanges some of the bits (genes) of the two chromosomes, whereas mutation inverts any bit(s) of the chromosome depending on a probability of mutation. Thus a 0 may be changed to a 1 or vice versa.

Then again step 2 is followed until the condition for ending the algorithm is reached.

In this research, the properties of genetic algorithm is as follow

a) Representation of the population
It is known that the desired solution is the optimum diameter and pressure distribution. In population, the solutions are arranged like binary code in a matrix as follows.

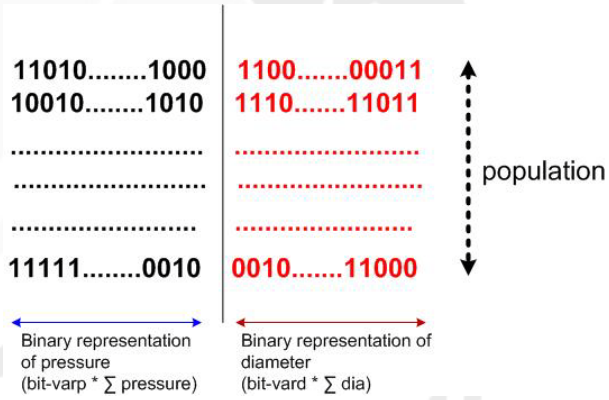


Figure 3. Binary representation for the pressure and diameter in population

bit_varp and *bit_vard* are the number of bits which represent the pressure and diameter index, respectively. *ΣPress* is the number of nodes that will be determined pressure, and *ΣDia* is the number of pipe segments which will be optimized in diameter. The binary representation of diameter represents diameter specifications which available on the market (ANSI / ASME) with a diameter range from 3 to 16 inch (64 kinds of diameter).

b) Operator selection, crossover, and mutation.
Operator selection, crossover, and mutation performed for each part of the pressure and diameter. This operator is based on any probability value.

c) Fitness function.
Fitness function is formulated as follows.

$$\min F(x) = \frac{1}{1+\|f(x)\|}, \quad wh\ re\ \|f(x)\| = \sqrt{f_1^2(x) + f_2^2(x) + \ldots + f_n^2(x)} \quad (11)$$

with *f* is a nonlinear equation of continuity that has been defined previously (Eq. (3) and (4) for each node).

AG is used to estimate the solution of the equation system from which f (x) = 0, so the best fitness value is when $F = 0$.

After the best individual which has the lowest fitness in step "c" has been gotten, the total cost will be calculated. Then, the result is sorted in ascending order. The lowest of total cost will be saved as the best individual on population for the next iteration. This process will be conducted until the maximum generation.

The last process of computation is the Newton's method. This method is used to the last checking whether the result had been gotten from genetic algorithm convergent and balanced perfectly. The detail of the process could be found in the previous paper [5].

3. CASE STUDY

In this paper, a developed model tested in the case of the X region with a network schematically as shown below.

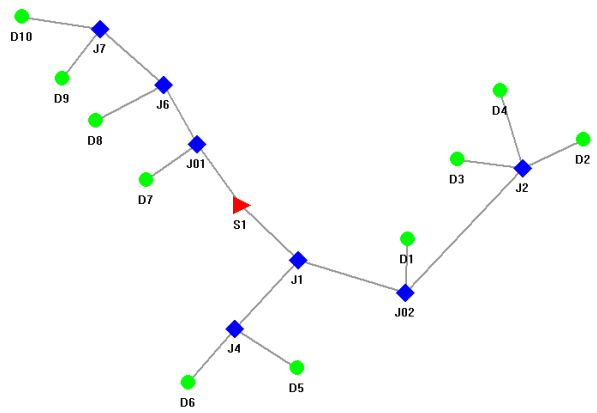


Figure 1. Schematic of gas distribution pipeline

Node S1 is the point of suppliers, while nodes D1, D2, D3, D4, D5, D6, D7, D8, D9, D10 are the point of demand, and nodes J01, J02, J1, J2, J4, J6, J7 are the points of junction. Pressure data and flow rate can be seen in the table below.

The optimum size of the pipe diameter at each segment in the network will be determined, also the pressures at junction J01, J6, J7, J1, J4, J02, J2 and 10 will be determined. The pressures on the other nodes have been determined (according to the contract). Pressure data and flow rate at each node can be seen in the following table.

Table 1. Data inputs for pressure and flow rate on the network

No.	Node	Pressure (*psia*)	Flow Rate (*MMscfd*)
1	S1	255	45.882
2	J01	Unknown	0
3	D7	247	-1.235
4	J6	Unknown	0
5	D8	240	-0.832
6	J7	Unknown	0
7	D9	221	-2.369
8	D10	196	-16.209
9	J1	Unknown	0
10	J4	Unknown	0
11	D5	250	-1.644
12	D6	245	-1.273
13	J02	Unknown	0
14	D1	245	-6.284
15	J2	Unknown	0
16	D2	228	-8.502
17	D3	225	-3.542
18	D4	230	-3.994

While the data of pipe length at each segment is as follows.

Table 2. Length of each segment of data on the network pipe

From node	To node	Distance (mile)
S1	J01	1.0501
J01	D7	0.1242
J01	J6	1.6606
J6	D8	0.15152
J6	J7	2.3306
J7	D9	0.3126
J7	D10	4.7249
S1	J1	0.57778
J1	J4	1.3823
J4	D5	0.28243
J4	D6	4.5863
J1	J02	0.65143
J02	D1	0.1242
J02	J2	2.3176
J2	D2	0.1242
J2	D4	1.1177
J2	D3	0.5589

Data inputs for economics factor are as follows.

- Pipe cost = US$ 2500/*ton*
- Coating cost = US$ 10/*meter*

- Installation cost = US$ 20/*inch*/*meter*

While input for Genetic Algorithm is as follows.

- Population = 50
- Persen_crossover = 90%
- Persen_mutation = 1%
- Max_generation = 3.000.

From the simulation results obtained as follows.

Table 3. Result of pressure distribution

No.	Node	Pressure (*psia*)
1	J01	246.6
2	J6	234.2
3	J7	219.3
4	J1	249.6
5	J4	248.5
6	J02	244.2
7	J2	234.3

Table 4. Result of the optimum pipe diameter at each segment

From Node	To Node	Inside Diameter (inch)	Wall Thickness (inch)
S1	J01	7.9	0.344
J01	D7	6,065	0.28
J01	J6	7.9	0.344
J6	D8	4.062	0.219
J6	J7	8.125	0.25
J7	D9	6.249	0.188
J7	D10	8.125	0.25
S1	J1	8.249	0.188
J1	J4	6.065	0.28
J4	D5	6.001	0.312
J4	D6	4.062	0.219
J1	J02	8.125	0.25
J02	D1	6.187	0.219
J02	J2	8.249	0.188
J2	D2	4.124	0.188
J2	D4	6.065	0.28
J2	D3	4.124	0.188

From the optimum diameter above, the economies cost obtained as follows.

Table 5. Result of the economies cost

No.	Item of cost	Cost (US$)
1	Investment	2,965,132.42
2	Coating	396,032.68
3	Installation	5,733,666.24
4	Operation	1,344,466.04
Total cost		10,439,297.38

After getting the result of diamater optimization using genetic algorithm, to get more satisfying, we should check using balancing3 system software to calculate pressure distribution on each node. Balancing system software what we have is using combination between genetic algorithm and newton's method [5]. After that, we should compare between the pressure given by user as input data on each demands and the result of balancing system software. We should run again if the result of comparation isn't good enough.

The result of pressure diameter of this case study is as follow.

Table 6. Result of the pressure distribution

No	Node Name	Pressure	Rate
		(Psia)	(MMscfd)
1	J01	246.569	0
2	J6	234.162	0
3	J7	219.309	0
4	J1	249.578	0
5	J02	244.221	0
6	J2	234.324	0
7	J4	248.506	0
8	S1	255	45.884
9	D10	193.628	-16.209
10	D9	219.147	-2.369
11	D8	234.076	-0.832
12	D7	246.549	-1.235
13	D6	243.064	-1.273
14	D5	248.426	-1.644
15	D1	243.851	-6.284
16	D3	229.99	-3.542
17	D4	232.661	-3.994
18	D2	229.439	-8.502

The main interface of the software which has been resulted based on the model is as follow.

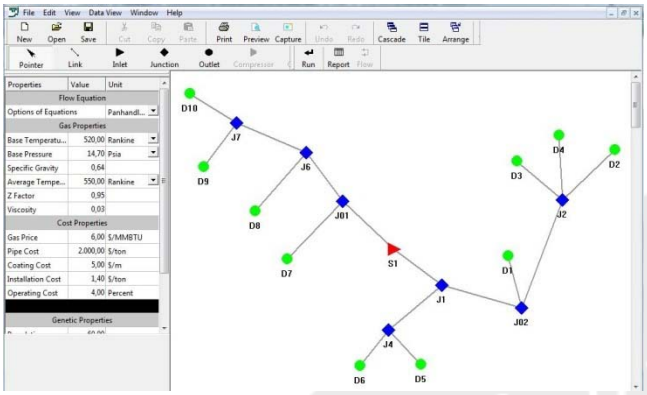


Figure 2. The main interface of the optimization software

4. CONCLUSION

Simple Genetic Algorithm can be helpful in finding an optimal inside diameter. The optimal inside diameter gives a minimum cost by considering the technical specifications (the pressure given by user). To give more satisfying result, the optimal inside diameter should be checked by software for calculating pressure distribution (balancing system software).

5. ACKNOWLEDGMENTS

Our thanks to RC-OPPINET–ITB Team for everything to make the research could be done.

6. REFERENCES

[1]. Stoner, M.A. 1969. Steady-State Analysis of Gas Production, Transmission and Distribution System. paper SPE 2554 presented at the SPE 44th Annual Fall Meeting, Denver, Colo. (Sept. 28-Oct. 1, 1969).

[2]. Goldberg D.E.. 1989. *Genetic Algorithm.* Addison-Wesley Publ. Co., Inc.

[3]. Sidarto K.A., Saiman dan N. Rohani. 2004. Menentukan Akar Sistem Persamaan Tak Linier dengan Memanfaatkan Algoritma Genetika yang Dilengkapi Clearing Procedure dari Petrowski. In *Proceedings of Konferensi Nasional Matematika XII* (Denpasar, Bali, 23-27 Juli, 2004).

[4]. American Petroleum Institute. 1980. *API Specification for Line Pipe*. American Petroleum Institute. Washington, D.C.

[5] Sidarto, K. A., Mucharam, L., Riza, L.S., Mubassiran, Rohani, N., Soplan, S. 2005. Implementation of Genetic Algorithm to Improve Convergente of Newton's Method in Predicting Pressure Distribution in Complex Gas Pipeline Network Sistem Case Study: Off-take Station ST-WLHR Indonesia. In *Proceedings of Seminar nasional soft computing, intelligent sistems and information technology/SIIT* (28-29 July 2005).

[6]. Flanigan O. 1972. Constrained Derivatives in Natural Gas Pipeline System Optimization. *Journal of Petroleum Technology* (May 1972), pp. 549 – 556.

[7]. Ikoku C.U. 1984. *Natural Gas Production Engineering*. John Wiley & Sons, New York.

[8]. Agarwal V. Solving Transcendental Equations Using Genetic Algorithm. http://www.geocities.com/mumukshu/gatrans.html.

Hybrid Genetic Algorithm for Solving Strimko Puzzle

Samuel Lukas
Informatics Department
UPH Tower, Lippo Karawaci
Indonesia
slukas@uph.edu

Arnold Aribowo
Computer Engineering Department
UPH Tower, Lippo Karawaci
Indonesia
Arnold.aribowo@staff.uph.edu

James Nagajaya Dyalim
Informatics Department
UPH Tower, Lippo Karawaci
Indonesia
Jsmx1llime@yahoo.com

ABSTRACT

Strimko is a challenging logic numbers puzzle. It has simple rules and could become a model of complicated problem. In order to find the solution of the Strimko problem, research is conducted and software is developed through this research.

The developed software is built based on heuristic search and genetic algorithm called hybridgenetic algorithm. If the heuristic search can not solve the problem, hybrid genetic algorithm will be executed. A single gene in a chromosome represents the index number of the empty cell, while sequence of genes in a chromosome represents sequence of empty cells to be solved.

Testing of various numbers of grids in Strimko problem is done after the software is built. The result of testing shows that the range of the software's finishing time is increased as the number of the grid in the Strimko problem increases. The range of the software's finishing time are 1 second for 4 x 4 grids, 1 – 7 seconds for 5 x 5 grids, 1 – 37 seconds for 6 x 6 grids, and 2 – 48 seconds for 7 x 7 grids.

Keywords

Heuristic search, Genetic algorithms, Strimko

1. INTRODUCTION

Strimko is a logic puzzle with numbers, based on a well-known idea of Latin squares described in the 18th century by a famous Swiss mathematician and physicist Leonhard Euler. That puzzle has a simple rule, but can involve one's complex logic to solve. The puzzle becomes even harder when the player have to choose more than one number possibility that can fill the grid in the Strimko, which decreases the chance of solving this puzzle completely. Based on that fact, an efficient method using the aid of the computer is necessary.

Strimko is a logic numbers puzzle and classified as a logic grid numbers. The concept of the puzzle is very simple. The Strimko puzzle problem with grids 5 x 5 is illustrated in Figure 1. It has three basic elements : rows, columns, and streams. Each element consists of five circles or grids. The content of each grid is a number, a member of {1,2,3,4 5 }.

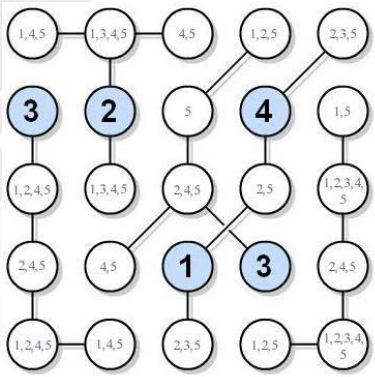


Figure 1. Strimko puzzle problem with grids 5 x 5

The problem is only some of the grids are known, The goal is to fill others empty grids so that the three basic elements containing the whole set of specified numbers. Solving Strimko puzzle manually may need a lot of time. In this research, a software is created by using two logic rules in the heuristic search and hybrid genetic algorithm to solve the problem.

2. HEURISTIC SEARCH

Heuristic search is a search strategy to solve a searching problem. It guides the user in a searching process in the highest success chance and throw off the unnecessary process. The heuristic search that is used in this research are open single and hidden single. A strimko puzzle with *n x n* grid, there are *n* possible number to become a content of a blank grid. By applying heuristic search on a blank grid, the possible number to fill the blank grid can be reduced.

In figure 1, the possible number to fill a grid in the second row and the third column can be reduced into a single number that is 5. Therefore, the content of that grid will be 5. This rules is called open single logic rule [2]. This is a basic rule in the logic puzzle. However, In Figure 2, open single rule can not be applied. If the first stream is defined as grids, {(1,1),(2,1),(2,2),(3,1),(4,1)}, then the content of grid (3,1) can be either 4 or 5. However, 4 can also be used in grid (2,1) or (2,2). But 5 can not be filled in other grids except in grid (3,1). This rule is hidden single rule [3].

There are several logic rules that can be applied for solving strimko. However, they are quite complicated to be implemented in programming. This paper only uses the two rules above. The other rules will be represented with genetic algorithms that we called as hybrid genetic algorithm. This hybrid genetic algorithm is a modification of genetic algorithm by combining with heuristics search above to produce a better solution in term of time.

Encoding, crossover and mutation that are used in this paper are respectively permutation encoding, cycle crossover and reciprocal exchange mutation.

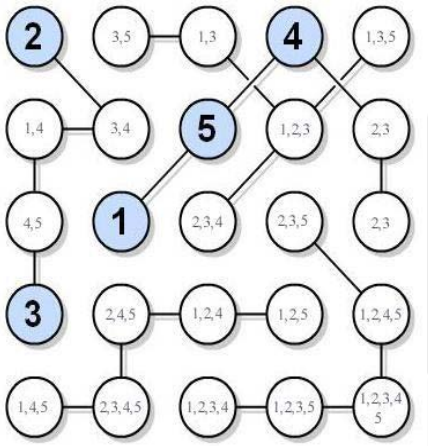


Figure 2. Hidden single rule in grids 5 x 5

3. DESIGN AND ANALYSIS

Given a strimko puzzle problem *n x n* grids. It means that there are *n* rows and *n* columns with *n* streams. There are several terminologies to be introduced. Firstly, each grid located in i^{th} row and j^{th} column is numbered with (1)

$$g(k) = (i-1)*n + j.\ k \in \{1,2,3,...,n^2\} \quad \text{.....(1)}$$

The content of $g(k)$ is noted as $v(k)$, $v(k) \in \{1,2,3,...,n\}$. Secondly, two matrices are created, called M matrix and S matrix. The M matrix is to allocate the number of each grid related to the row and the column whereas S matrix to that of the streams. m_{ij} and s_{ij} are elements of *M* and *S* matrix located in i^{th} row and j^{th} column respectively. m_{ij} represents the number of the grid $g(k)$. Where as s_{ij} represents the number of the grid of i^{th} stream and j^{th} sequance of the stream. This representation can be seen in figure 3 and Table 1.

The problem is what is the value of $v(m_{ij})$ and $v(s_{ij})$ if value of some certain grids are known so that they satisfy equations

$$\sum_{j=1}^{n} v(m_{ij}) = \sum_{x=1}^{n} x,\ i = 1,2,...,n \quad \text{.....(2)}$$

$$\sum_{i=1}^{n} v(m_{ij}) = \sum_{x=1}^{n} x,\quad j = 1,2,...,n \quad \text{.....(3)}$$

$$\sum_{j=1}^{n} v(s_{ij}) = \sum_{x=1}^{n} x,\quad i = 1,2,...,n \quad \text{.....(4)}$$

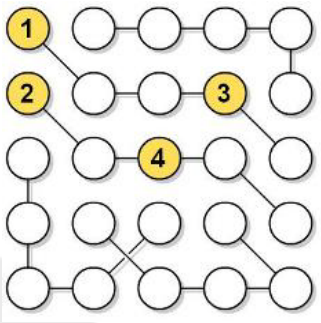


Figure 3. A strimko problem grid 5 x 5

Tabel 1: M and S matrix for Figure 3.

1	2	3	4	5
6	7	8	9	10
11	12	13	14	15
16	17	18	19	20
21	22	23	24	25

M Matrix of A

1	7	8	9	15
2	3	4	5	10
11	16	18	21	22
6	12	13	14	20
17	19	23	24	25

S Matrix of A

The heuristics algorithm to solve the problem is stated below :

1. Generating all possible value to be put in each blank k^{th} grid without breaking all strimko rules, $P(k)$, by applying the mathematic model in Equation (5).

$$P(m_{ij}) = \{1,2,...,n\} - \bigcup_{y=1}^{n} v(m_{iy}) - \bigcup_{x=1}^{n} v(m_{xj}) - \bigcup_{y=1}^{n} v(s_{iy}) \quad (5)$$

2. Applying heuristic search by open single logic rule for all blank grid.
3. Repeat step 1 to 2 until open single logic rule for all blank grid can not be applied.
4. Applying heuristic search by hidden single logic rule for all blank grid.
5. Repeat step 1 to 3 until hidden single logic rule for all blank grid can not be applied.
6. If the problem has not been solved then hybrid genetic algorithms will be executed.

Hybrid genetic algorithm is a modified genetic algorithm in which the content of each gene in a chromosome is not generated randomly but heuristically. A chromosome, labeled as string *C*, is segmented into *n – a* segments, where *a* is the total of the stream that has not an empty grid. The segment is ordered by the stream that has least empty grid. In each segment contains genes which are order by the number of the empty grid in that stream. *C(x)* is a x^{th} gene in a chromosome that associate with a stream and an empty grid at that stream. If that empty grid, located in i^{th} row and j^{th} column, is *g(k) according to (1)* then $C(x) = g(k)$.

The process of valuing x^{th} gene in a chromosome, that is $v(k)$, is started by determining the value of the first gene randomly but under all the possible values of that gene. Then, it is continuing by

valuing other genes which are relating with that first gene by using the heuristic search. If the heuristic search cannot determine other values of the genes, the system will randomly determine the value of the most left gene that has no value. This process is done until there is no value available to a gene then the value of that gene and others unprocessed gene are set to 0. $w(x) \in \{0,1\}$ is a sign to indicate the x^{th} gene has not or has a value and if the length of a chromosome is z then the fitness of that chromosome is (5)

$$fitness = \frac{z - \sum_{x=1}^{z} w(x)}{z} \quad \ldots\ldots (6)$$

If fitness value of a chromosome is one, it means that the chromosome is the best chromosome and as a solution of the problem.

A certain number of chromosomes is generated in the first population. If there is no chromosome has fitness equal to one then the system will generate the next population by doing cycle crossover and reciprocal exchange mutation. The system will continue doing hybrid genetic algorithm until a chromosome with fitness value 1 is found.

4. EXPERIMENT RESULTS

Three experiments are exercised to analyze the software in solving the strimko problem. The first experiment is analyzing the genetic algorithm's parameters. There are three genetics algorithm parameters, population, probability of crossover and also mutation. Given a specific strimko problem, we would like to know whether there is a best set of genetic algorithm parameter so that the time needed to solve the problem is the fastest. Each set of parameter is trying three times and each parameter is changed 10 times from 10 to 100. Therefore 3000 experiments are conducted. The result of these experiments is tabulated in Table 2. From the table can be seen that there is no best set of parameters for each case. But, it is proven that the system work well.

The second experiment is to determine the fastest time needed to solve the defined problem. From the first experiment in genetic algorithm's parameters experiment, it showed that for 10% mutation, the best population was 20 and the percentage of cross over was 50%. From this set of parameter, twenty experiments were conducted to know the average time needed to solve the problem. Other conditions are also investigated and the result is tabulated in Table 3.

The results of further investigation from the set of combinations were the average time to solve the problem was six seconds with standard deviation 7.9 second, while the worst was 7 seconds with standard deviation 7.4 second.

The third experiment is to find out the relation between the average time needed to solve the problem and the size of the problem. For the size of grid 5 x 5, 6 x 6 and 7 x 7, software can solve the problem with maximum time 7, 37 and 48 second respectively.

Table 2. Genetic algorithm's parameters experiment

Probability of Mutation	First Experiment		Second experiment		Third experiment	
	Best population	Best crossover	Best population	Best crossover	Best population	Best crossover
10%	20	50%	80	50%	70	90%
20%	100	10%	60	80%	30	40%
30%	90	30%	20	50%	70	80%
40%	40	10%	90	40%	10	100
50%	10	40%	10	80%	90	90%
60%	70	40%	40	90%	70	30%
70%	90	10%	50	70%	50	100
80%	60	40%	50	30%	30	100
90%	50	60%	80	70%	70	80%
100	90	10%	20	70%	90	30%

Table 3. Time needed to solve a strimko puzzle base on Table 1

Mutation probability	Time (second) needed in first experiment	Time (second) needed in second experiment	Time (second) needed in third experiment
10%	11	10	11
20%	8	11	12
30%	12	15	11
40%	11	16	11
50%	9	15	10
60%	11	13	15
70%	11	16	13
80%	11	12	15
90%	13	15	15
100%	14	11	14

5. CONCLUSIONS

There are 2 points that can be concluded from this research.

1. Strimko puzzle can be solved using heuristic search or hybrid genetic algorithm.
2. Solving time range for Strimko problem is increased as the number of the grids in Strimko problem increases.

6. REFERENCES

[1] Website Strimko, About, http://www.Strimko.com/about.htm

[2] Strictly Sudoku, Open Single, http://www.strictlysudoku.com/open-single.php

[3] Sudoku Hints, Solving Sudoku, http://www.angusj.com/sudoku/hints.php

Optimal Design of Hydrogen Based Stand-Alone Wind/Microhydro System Using Genetic Algorithm

Soedibyo
Sepuluh Nopember Institute of Technology (ITS)
Sukolilo surabaya
dibyo_55@yahoo.com

Heri Suryoatmojo
Sepuluh Nopember Institute of Technology (ITS)
Sukolilo surabaya
heri_atmojo@yahoo.co.in

Imam Robandi
Sepuluh Nopember Institute of Technology (ITS)
Sukolilo surabaya
robandi@ee.its.ac.id

Mochamad Ashari
Sepuluh Nopember Institute of Technology (ITS)
Sukolilo Surabaya
ashari@ee.its.ac.id

Takashi Hiyama
Kumamoto university
Kurokami 3-3-91, Japan

ABSTRACT

The main targets of optimization problem in the stand alone hybrid generation system is to satisfy the load demand with high reliability and respect to minimum annual cost of system. This paper utilizes Genetic Algorithm (GA) method to determine the optimal capacities of hydrogen, wind turbines and microhydro unit according to the minimum cost objective functions that relate to these two factors. In this study, the cost objective function includes the annual capital cost (ACC) and annual custumer damage cost (ADC). The proposed method has been tasted in the hybrid power generation system located in Sirengge village in Central Java of Indonesia. Simulation results show that the optimum configuration can be achieved using 240 ton of hydrogen tanks, 0 wind turbines and 3 x 2.1 MW of microhydro unit respectively.

1. INTRODUCTION

Nowadays, renewable energy has been explored to meet the load demand. Utilization of renewable energy is able to secure long-term sustainable energy supply, and reduce local and global atmospheric emissions [1]. Microhydro (Hyd) and Wind Generation (WT) units are become the promising technologies for supplying the load demand in remote and isolated area. However, there are several weakness faced by such resources. One of the weakness is the power generated by wind and microhydro energy is influenced by the weather conditions. The variations of power generated by these sources may not match with the time distribution of demand. In addition, the intermittent power from wind and microhydro power may result in serious reliability concerns in both design and opreation of microhydro and wind turbines system. For simplicity, to overcome the reliability problem, over sizing maybe can be applied. However, installing the components improperly will increase overall cost system.

Actually, there are several alternative ways to prevent the shortage power from these power. A back-up unit can be considered as a power supply whenever the isufficiency power is occurred. For instance, diesel generator is one of the alternative back-up power. However, the operational cost of diesel generator is considerably high, also utilization of diesel generator is not the good option due to the environmental concern. Menwhile, battery storage also can be considered for the back-up unit. However, the operational and maintenance procedures of battery are complicated. The last back-up unit goes to utilization of fuel cell equipped with electrolyzer and hydrogen tank.

The most important challenge in design of such systems is reliable supply of demand under varying weather conditions, considering operation and investment costs of the components. Hence, the goal is to find the optimal design of a hybrid power generation system for reliable and economical supply of the load [2]. Several method has been done by another researcher, many papers offer a variety of methods to find the optimal design of hybrid wind turbine and photovoltaic generating systems [2-3]. In [2] and [4] Genetic Algorithm (GA) finds optimal sizes of the hybrid system components. In some leter research, PSO is successfully implemented for optimal sizing of hybrid stand-alone power system, assuming continuous and reliable supply of the load [3]. However, none of them working with the microhydro system.

This paper proposes the method to find the optimal design of hybrid power generation system consists of microhydro, wind turbine and fuel cell in the system. The target is to find the optimal size of components respect with minimum total annual cost (ACS). In this way, genetic algorithm is utilized to minimize cost of the system over its 20 years of operation, subject to reliability constrain. Wind speed and stream flow data are available for Sirenggi village in Banyumas, Indonesia and system costs include Annualized Capital Cost (ACC), as well as costumers' dissatisfaction cost. Next section briefly describes the hybrid system model.

2. SYSTEM CONFIGURATION

Block diagram of a hybrid Microhydro and Wind Turbines system is depicted in Fig.1. The typical daily load demand can be seen in Fig. 2. The hybrid system consists of 2 types of power generator; wind turbines unit and microhydro unit connected to the load system through the inverter. The storage system consists of

electrolyzer, hydrogen storage tank and fuel cell required to store all excess power. Detailed component model and their specification, used in this study will be explained in the following sections.

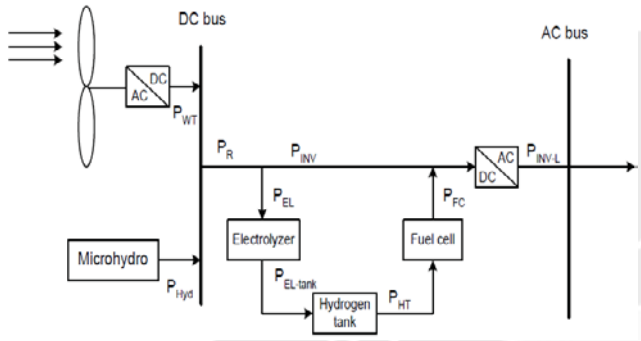


Figure 1. System configuration

Table 1. Specification of wind turbines

Rated power	750 kW
Cut-in speed	3 m/s
Rated speed	12.5 m/s
Cut-out speed	25 m/s
Swept area of rotor	1,964 m^2
Efficiency	31.92%

2.1. Wind Turbine Generator

The output power of each wind turbine (W) unit is based on the rated capacity and the specification given by the manufacture. In this study, UGE wind turbine (UGE 750H) is considered as a power generator. It has a rated capacity of 750 kW and provides 780 V DC at the output. The output power from UGE 750H can be described by the following equation.

$$P_W(t) = \begin{cases} 0 & \text{if } v_t(t) < v_c \\ \frac{1}{2}.\rho.A.v^3(t).\eta_{wt} & \text{if } v_c \le v_t \le v_r \\ P_{rated} & \text{if } v_r \le v_t \le v_f \\ 0 & \text{if } v_t > v_f \end{cases} \quad (1)$$

Where; ρ is air density kg/m^3. A is swept area of rotor m^2, t is wind speed (m/s), η_{wt} is efficiency of WTs, V_c is cut-in speed, v_r is rated speed, v_f is furling speed and P_{rated} is rated power of WTs. The specification of WTs as shown in Table 1.

2.2. Microhydro Output Power

The electrical power generated by the hydro turbine can be determined using the following equation [5]

$$P_{hyd} = \eta_{hyd}.\rho_{hyd}.h_{net}.Q_t \quad (2)$$

Where, H_{net} is the effective head, the actual vertical drop minus this head loss. It can be calculated using the following equation [5]:

$$h_{net} = h_{\_}(1 - f_h) \quad (3)$$

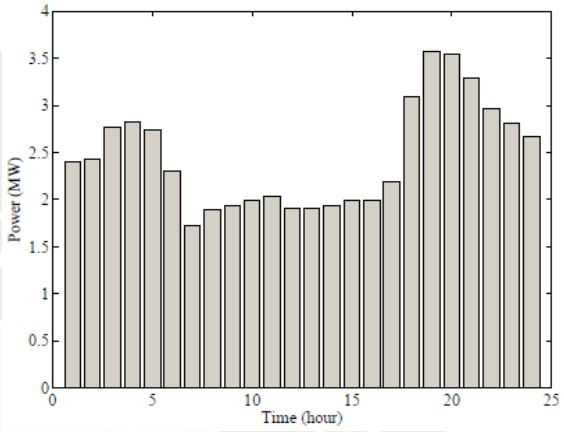


Figure 2. Load demand

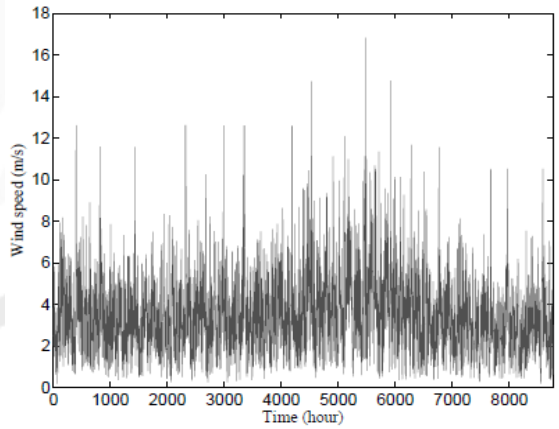


Figure 3. Wind speed data

Meanwhile, Q_t is the hydro turbine flow rate, the amaout of water flowing through the hydro turbine. It can be calculated using the following equation [5].

$$Q_t(t) = \begin{cases} 0 & \text{if } Q_{Av}(t) < Q_{min} \\ Q_{Av}(t) & \text{if } Q_{min} < Q_{Av}(t) < Q_{max} \\ Q_{max} & \text{if} Q_{Av}(t) > Q_{max} \end{cases} \quad (4)$$

Where; Q_{Av} is the flow rate available to the hydro turbine (m^3/s), Q_{min} is the minimum flow rate of the hydro turbine (m^3/s), Q_{max} is the maximum flow rate of the turbine (m^3/s) [5].

Q_{min} is the minimum flow rate, the minimum allowable flow rate through the hydro turbine, it is assumed that the hydro turbine can operated only if the available stream flow is equal to or exceeds this minimum value. It can be calculated using the following equation [5]:

$$Q_{min}(t) = W_{min} \cdot Q_D \quad (5)$$

Q_{max} is the maximum acceptable flow rate through the hydro turbine, expressed as a percentage of the turbine's design flow rate [5]. This simulation uses this input to calculate the maximum flow rate trough the hydro turbine, and hence the actual flow rate through the hydro turbine. The flow rate profile can be seen in Fig. 4.

$$Q_{max}(t) = W_{max} \cdot Q_D \quad (6)$$

2.3. Electrolyzer

Basically, electrolyzer work based on the water electroysis. A direct current is passed between two electrodes then submerged in water and decomposes into hydrogen and oxygen. Then, the amount of hydrogen can be collected from the anode side. Usually, the hydrogen produces by the electrolyzers at a pressure around 30 bars. Also, the reactant pressures within a Proton Exchange Membrane Fuel Cell (PEMFC) are around 1.2 bar. For assumption, the electrolyzer is directly connected to the hydrogen tank. Transferred power from electrolyzer to hydrogen tank can be defined as follows [3]:

$$P_{EL-tank} = P_{EL} \times \eta_{el} \quad (7)$$

Where, η_{el} is the efficiency of electrolyzer.

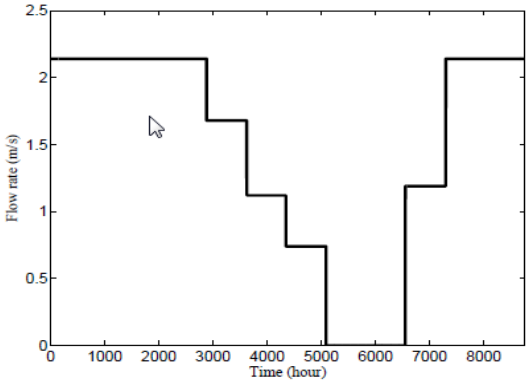


Figure 4. Flow rate of microhydro turbine

Table 2. Economic parameter considered for system optimization

Nominal interest rate i' (%)	8.25
Inflation rate f (%)	8.17
Project lifetime (years)	20.00
Wind turbines lifetime (years)	20.00
Hydrogen tanks lifetime (years)	20.00
Electrolyzer lifetime (years)	20.00
Fuel cell lifetime (years)	20.00
Inverter lifetime(years)	20.00
Cost of hydrogen (US$/kW)	1,500.00
Cost of electrolyzer (US$)	2,000.00
Cost of fuel cell (US$)	3,000.00
Cost of microhydro of 2.1 MW (US$)	8,000,000.00
Cost of wind turbine of 750 kW (US$)	1,000,000.00
Cost of inverter (US$/kW)	800.00

2.4. Hydrogen Tank

The basic principle of energy stored in the hydrogen tanks is the same as in the battery banks. Every hour energy stored in the hydrogen tanks can be described by using the following equation [3]:

$$E_{tank}(t) = E_{tank}(t-1) + \left(P_{EL-tank} - \frac{P_{HT}(t)}{\eta_{storage}}\right) \quad (8)$$

Where, P_{HT} is the power tranfered to the fuel cell. Here, it is assumed the hydrogen tanks efficiency is 98%. Meanwhile, the mass of stored hydrogen, at any time step t, is calculated as follows [3]:

$$m_{tank}(t) = \left(\frac{E_{tank}(t)}{HHV_{H_2}}\right) \quad (9)$$

Where, the Higher Heating Value (HHV) of hydrogen is equal to 39.7 kWh/kg. The energy stored in the hydrogen tanks can not exceed the constraint as follows [3]:

$$E_{tankmin} \leq E_{tank}(t) \leq E_{tankmax} \quad (10)$$

2.5. Fuel Cell

Fuel cells are electrochemical devices to convert the chemical energy of a reaction directly in to electrical energy. The output power produced by fuel cell can be determined by multiplying its input power and efficiency (η_{FC}). In this case the efficiency of fuel cell is assumed to be 50% [3].

$$P_{FC-inv} = P_{tank-FC} \times \eta_{FC} \quad (11)$$

2.6. Inverter

Inverter is electrical devices to convert electrical power from DC into AC form at the desired frequency of the load [3].

$$P_{INV-L} = (P_{FC} + P_{INV}) \cdot \eta_{inv} \quad (12)$$

Where, η_{inv} is inverter efficiency.

3. RELIABILITY AND COST

In this study, the objective function is the annual cost of system (ACS). The ACS model is suitable to find the best benchmark of cost analysis. Annual cost of system convers the annual capital cost (ACC) and annual damage cost (ADC). The components to be considered are wind turbine, microhydro, electrolyzer, hydrogen tank, fuel cell and inverter. ACS is calculated in the following equation [1]:

$$ACS = ACC + ADC \quad (13)$$

The annual capital cost of each units is calculated as follows [1]:

$$ACC = C_{cap} CRF_{(i,y)} \quad (14)$$

Where; C_{cap} is the capital cost of each component in US$, y is the project lifetime in year. CRF is capital recovery factor, a ratio to

calculate the present value of a series of equal annual cash flows. This factor is calculated as follows [1]:

$$CRF = \frac{i(1+i)^y}{(1+i)^y - 1} \quad (15)$$

Where; i is the annual real interest rate. The annual real interest rate includes the nominal interest and annual inflation rates. This rate is calculated as follows [1]:

$$i = \frac{(i' - f)}{(1+f)} \quad (16)$$

Cost of electricity shortages (ADC) can be estimated in different concept. The values of ADC usually can be considered in the range of 5 – 40 US$/kWh for industrial users and 2 – 12 US$/kWh for domestic users. In this simulation, cost of custumer's dissatisfaction, caused by loss of load is assumed to be 5.6 US$/kWh [3].

3.1. Optimal Operation Strategy

Basically, the optimal operation strategy contains of power flow simulation in order to supply the load demand. Here, the basic concept of strategy of system operation can be explain as follows.

If $P_R(t) = P_{load}(t)/\eta_{inv}$, in this case the entire power generated by the renewable sources is supplied to the load through the inverter.

If $P_R(t) > P_{load}(t)/\eta_{inv}$, the surplus power is consumed by the electrolyzer. However, if the excess power exceeds the rated capacity of electrolyzer, then the excess power will be dumped.

If $P_R(t) < P_{load}(t)/\eta_{inv}$, the insufficiency power will be supplied by the fuel cell. However, if the insufficieny power exceeds fuel cell capacity or the stored hydrogen cannot afford that, some fraction of the load must be shed. Then, the loss of load is occurred.

3.2. Optimization Procedure Using Genetic Algorithm

The simulation method utilizes genetic algorithm (GA) to determine the optimal sizing of the hybrid system. The concept of GA is defferent from traditional searh and optimization method used to solve the engineering problems. The basic idea of GA is taken from genetic process in biology that used artificially to build searh algorithms. This technique is introduced to find the optimal solution based on natural selection. The main objective of the proposed method is to find the optimum size of the hydrogen tanks, number of wind turbines and number of microhydro. To procced this study, the annual data of flow rate of river, wind speed and load demand are initially set as the input. Then, the size of hydrogen tanks, wind turbines and microhydro are randomly chosen to be GA chromosomes. Each cromosome consists of three genes in form of $[N_{H_yT} / N_{WT} / N_{Hyd}]$;

Where N_{HyT} is the number of hydrogen tanks, N_{WT} is the number of wind turbines and N_{Hyd} is number of microhydro. After setting the initial population, the annual power supply simulation are performed.

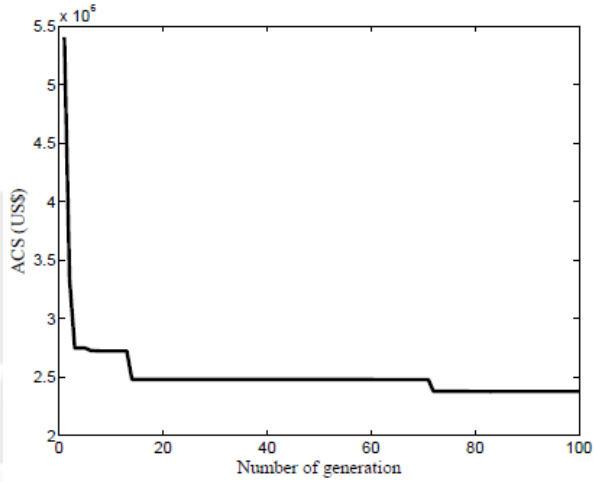


Figure 5. Optimization process using GA

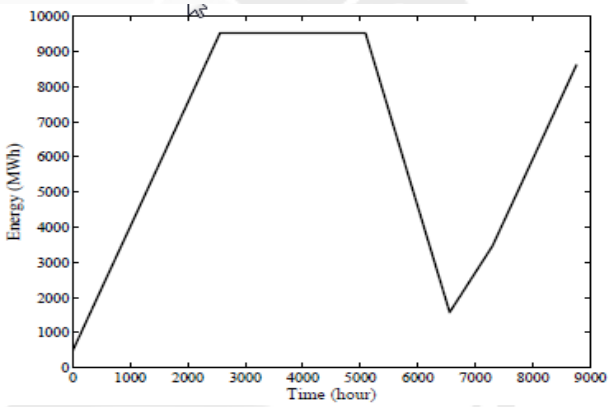


Figure 6. Level of energy stored in the hydrogen tanks

The simulation of annual power supply are repeated for each chromosome until it reaches the final generation as defined in the beginning of the simulation process. Each generation of the best chromosome is preserved and compared with the best chromosome obtained from the next generation. The best chromosome in the final generation is considered as the optimum parameter value of the hybrid system. In order to select the chromosomes subjected to the crossover and mutation for prcessing the next generation population, the roulette wheel method is considered as the selection process. In this simulation, the crossover and mutation probability are assumed as 0.75 and 0.015, respectively.

4. APPLICATION AND RESULTS

Genetic Algorithm based matlab m-file code has been developed to determine the optimal sizing of hydrogen-wind turbines-microhydro system in Sirengge (Central Java of Indonesia). The daily load profiles are represented by a sequence of powers which is constant over step time of one hour as shown in Fig. 2. In this simulation, the daily load profile is repeated within a year time based simulation. Hence, during one year, there are not differences between the day and the others day. The wind speed data can be seen in Fig. 3.

Meanwhile, the typical flow rate of turbines in Sirengge is shown in Fig. 4. Meanwhile, another data used for the optimization are shown in Table 2. In this simulation, GA parameters consists of 20 populations, and 100 maximum generations. Each chromosome consists of 3 genes which represent the size of hydrogen tanks, wind turbine and number microhydro. The values of crossover and mutation probability are 0.75 and 0.015, respectively. These values are determined by trial and error in order to find the optimal value quickly.

The convergence curves of the GA for the system under study is shown in Fig. 5. It can be seen that the optimal values can be obtained closed to 70 generations. Hence, 100 iterations can be considered as a fair termination criterion.

Table 3 shows the optimization result for the system under study. Here, the ACS value of proposed configuration is 2.3757 million US$. From the optimization process, the candidate of components used in this area consist of Hydrogen Tanks and Microhydro unit systems. It means that wind turbine is not selected to be installed in this area. From the Fig. 6, it can be explain that the possibility of power generated by microhydro system during the first month until the drought season is enough to supply the load demand. However, while the drought season the flow rate of river is very small and not enough to spin the hydro turbine. Hence, during the drought season the power stored in the hydrogen tanks used to supply the entire load demand. In this simulation it is assumed that the initial condition of energy stored in the hydrogen tanks is the minimum level.

In this model, the inverter capacity can be determined from the peak power demand and the efficiency of inverter. Meanwhile, the capacity of electrolizer is determined from the maximum power through the elctrolizer. Finally, the fuel cell capacity is determined from the maximum load demand minus maximum power generated by renewable sources.

Fig. 7 depicts the impact of interest rate to the value of ACS. Increasing the interest rate will influence the condition of annual payment to the bank. By using the optimal size of components the proposed configuration never has the electricity interruption therefore the value of ADC as representation of interruption the value is zero. It means that the proposed configuration has 100% reliability.

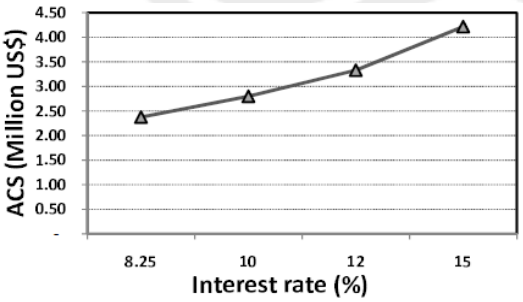


Figure 7. Impact of interest rate to the ACS value

Table 3. Optimization results

M_{Tank} (ton)	N_{WT}	N_{Hyd}	P_{EL} (MW)	P_{FC} (MW)	P_{INV} (MW)	ACS (US$ x 1e6)
240	0	3	4.2621	3.7534	3.7534	2.3757

5. CONCLUSION

Genetic algorithm method to find the optimal size of hydrogen, wind turbines and microhydro system has been presented in this paper. From the simulation results it can be observed that 240 ton of hydrogen tanks and 6.3 MW of microhydro unit are suitable to be used for electrification in the Sirengge village. Finally, this proposed method also could be used as optimization tool to find the optimal planning stage for electrification remote communities involving hydrogen-wind-microhydro hybrid system.

6. REFERENCES

[1] H. Suryoatmojo, A. A. Elbaset, T. Hiyama. 2009. Economic and Reliability Evaluation of Wind-Diesel-Battery System for Isolated Island Considering CO2 Emission. IEEJ Trans. PE, vol. 129, no. 8, pp. 1000-1008.

[2] H. Suryoatmojo, T. Hiyama , A. A. Elbaset, M. Ashari. 2009. Optimal Design of Wind-PV-Diesel-Battery System Using Genetic Algorithm. IEEJ Trans. PE, vol. 129, no. 3, pp. 413-420.

[3] A. Kashefi Kaviani, G. H. Riahy, S H. M. Kouhsari. 2009. Optimal Design of A Reliable Hydrogen-Based Stand-Alone Wind/PV Generating System Considering Component Outages. Renewable Energy, vol. 34, pp. 2380-2390.

[4] R. Dufo Lopez, J. L. Bernal-Agustin. 2005. Design and Control Strategies of PV-Diesel Systems Using Genetic Algorithms, Solar Energy, vol. 79, pp. 33-46.

[5] www.nrel.gov/homer/

[6] Hongxing Yang a, Wei Zhou a, Lin Lu a, Zhaohong Fang. 2007. Optimal sizing method for stand-alone hybrid solar–wind system with LPSP technology by using genetic algorithm, Solar Energy, vol. 82, 354–367.

Optimization of Steel Structure by Combining Evolutionary Algorithm and SAP2000

Mohammad Ghozi
Bhayangkara Surabaya University
114 Ahmad Yani st
Surabaya, Indonesia
031-8285602
mghozi2002@gmail.com

Pujo Aji
Sepuluh Nopember Institute of Technology
Kampus ITS Sukolilo Surabaya
031-5946094
pujo@ce.its.ac.id

Priyo Suprobo
Sepuluh Nopember Institute of Technology
Kampus ITS Sukolilo Surabaya
031-5946094
priyo@ce.its.ac.id

ABSTRACT
Evolutionary algorithms have been shown to be very effective optimization tools for a number of engineering problems and specially practical optimization problems. The use of nonlinear finite element can assist greatly in achieving a safe design. However, commercially available finite element programs are not designed for optimization tools. Design feature which included in commercial structure analysis program usually used to check if the member pass the code or not. 'Home-written' structure analysis program can be designed to achieve this task, however it may suffer from serious drawbacks such as bugs, lack of user friendliness, lack of generality, and unproven reliability. A new approach is presented for the optimization of steel truss by combining evolutionary algorithms and commercial structure analysis program. The purpose of this approach is to create a program which combine SAP2000 structure analysis program with evolutionary Algorithm for optimizing truss structure.
Strategies are discussed to model the chromosome and to code evolutionary to handle such constraints. Member grouping is created for reducing the problem size and implementing move-limit concepts for reducing the search space. The implemented process is tested on 10 members of 2D steel structure, 25 members of 3D steel structures and 36 members of 3D steel structure, where the results are compared with previous researches. Because of data size and number population, parallel computing method used in this paper. It is concluded that this method can serve as a very useful tool in engineering design and optimization.

Keywords
Optimization, evolutionary algorithm, structure analysis program, steel structure, SAP2000.

1. INTRODUCTION
The use of finite element method (FEM) which used by commercial FEM program can assist in achieving a safe design. However, commercially finite element programs are not designed for optimization. Design feature which included in commercial structure analysis program usually used to check if the member and applied inner force pass the corresponded code or not. 'Home-written' structure analysis program can be designed to achieve this task, however it may suffer from serious drawbacks such as bugs, lack of user friendliness, lack of generality, and unproven reliability [11].

The application of genetic and evolutionary computation to the automated design of structures has followed several avenues. The first is topology and shape optimization, in which the applications have included elastic truss structures subjected to static loading [12]. There have also been research efforts devoted to developing algorithms for optimized structure topologies to satisfy user-determined natural frequencies. The second major area of automated design using genetic algorithms has been their application for optimal member sizing for truss structures using linear elastic analysis with general stress criteria [12], or U.S. design specifications [1].

The final major application of genetic algorithms has been the automated design of steel frame structures. The vast majority of these efforts have been restricted to the optimized design of planar structures using linear elastic analysis. However, recent research efforts have begun to utilize genetic algorithms to guide the design of steel framed structures where the structural analysis includes nonlinear geometric behavior and nonlinear material behavior with semi rigid connections. Excellent methods were combining commercial FEM program with genetic algorithm to find shortest member's length [14] and combining commercial FEM program with iteration method to find required area of steel reinforced concrete plate [11].

2. THEORIES
In this section, SAP2000 Evolutionary algorithm and recent research will be described.

2.1 SAP2000 Structure Analysis Program
SAP2000 structure analysis program is well known as a Finite Element Analysis tool which already used for analyzing and modeling structure based on the relevant code such as AISC-LRFD99 [4].

SAP2000 could process or import the file input with extension MDB, XLS, TXT and SDB. SAP2000 also could export analysis result and design to files with extension XLS, TXT and SDB. After input file being opened, SAP2000 will run analysis, save result and design all members. In the output file, we can get required data such as frame stress, joint displacements and ratio of design criteria [3].

Design results data include the design stresses, stress ratios, effective lengths, optimal sections, area of reinforcing steel, and all other calculated quantities resulting from the design process. The main body of the form lists the design stress ratios obtained at various stations along the frame object for each design load combination. The SAP2000 automatically created code-specific design load combinations for this steel frame design.

Ratio in this paper is the comparation between actual inner force and member's capacity based on it's material property (e.g., cross sectional area, Inertia, etc.) based due to PMM method and AISC-LRFD99 Code. Ratio of design criteria has a value between 0 and 0,95. If member has ratio 0,1 it means the member is overstrength, if the ratio is 0,92 the member is efective, if ratio is 1,3 it means the member is overloaded.

In this paper, design criteria will be used as constraint and combined with EA for optimizing steel structure based on AISC-LRFD Code.

2.2 Evolutionary algorithm

Genetic algorithm (GA), a member of Evolutionary Algorithm (EA), is a population-based global search technique based on the Darwinian evolutionary theory [9].

The preliminary approach of GAs is Simple Genetic Algorithm (SGA, see a pseudocode in fig. 1). SGA guides the evolutionary search by a single population Pi. The size of Pi is denoted by SP. Individuals are encoded in a string scheme associated with one of the codes of the binary, integer, and real. In the evolutionary search, the promising individuals Pisel and Pi+1−sel are chosen from the population by a selection operation (roulette wheel, stochastic universal sampling, ranking, truncation, etc.). Then, the individuals chosen are applied to recombination and mutation operation (one point or multipoints crossover and mutation, uniform crossover, etc.). These evolutionary operations (mutation mut, crossover cr, and selection sel) are governed by their related evolutionary parameters Par (mutation and recombination probability rates, selection pressure, etc.). The population Pnew evolved by the application of these evolutionary operators is decoded. Then, the fitness values are computed by use of this population. The evolutionary search is executed to transmit (migration) the individuals (emigrant and immigrants)

SGA (P_i, NG, SP, F_i, Par_{sel}, Par_{mut}, Par_{cr})
If $[P_i] = []$, Initialize (P_i, SP, NDV)
for $i = 1$: NG
$[P_i^d] = P_i$
If required, $[P_i^d]$ = Decoding (P_i)
If $[F_i] = []$, $[F_i]$ = Fitness_Calculation (P_i^d)
$[P_{i-\text{sel}}]$ = Selection(P_i, F_i, Par_{sel})
$[P_{i+1-\text{sel}}]$ = Selection(P_i, F_i, Par_{sel})
$[P_{\text{new}}] = P_{\text{new}}$ U Crossover($P_{i-\text{sel}}$, $P_{i+1-sel}$, Par_{cr})
U Mutation($P_{i-\text{sel}}$, $P_{i+1-\text{sel}}$, Par_{mut})
$[P_i] = [P_{\text{new}}]$
end

Figure 1. Psoudocode of Simple Genetic Algorithm [10].

to the next populations until satisfying a predetermined stopping criteria (e.g., completion of a generation number NG). This is one big weakness in genetic algorithm that solving problem need a lot of generation and this can take time.

Evaluation function was a base step for selection process. In this phase, strings were converted to function parameter, evaluate the objective function and then convert the objective function to fitness. In general optimization problem which for maximize problem, the fitness is equal to objective function [8], but for minimize problem, the fitness is equal to

$$\text{Eval (Vk)} = 1/\text{ Ckmax} \qquad (1)$$

Where :

Eval (Vk) = objective function
Ckmax = fitness

To combine penalty function and objective function, penalty function is added with objective function, so the objective function equation :

$$eval_{(x)} = f_{(x)} + p_{(x)} \qquad (2)$$

where x is chromosome, f(x) is objective function and p(x) is penalty function.
For minimizing problem, penalty function are describe below :

$$p_{(x)} = 0 \quad \text{if x is preferred solution}$$

$$p_{(x)} > 0 \quad \text{if x is not preferred solution} \qquad (3)$$

3. METHOD

Problem in this paper is solved by studying behavior of FEM program, making GA which combine SAP2000-GA, making the parallel computing module, and running the program. Steps of this research described in figure below.

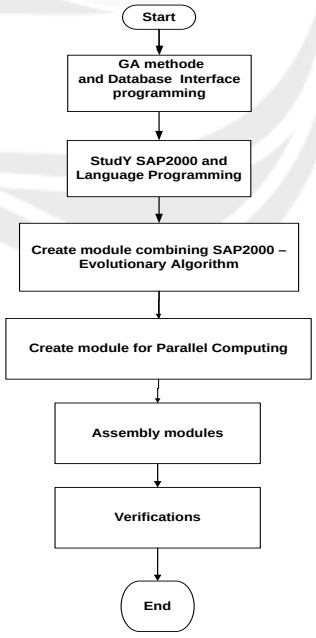


Figure 2. Flowchart to combine evolutionary algorithm and SAP2000.

The steps for optimizing steel structures are initialized by creating SAP2000 input file and then rechecking the geometry. Population is generated by creating more SAP2000 input files randomly. Randomly means the selected members defined randomly based on the provided materials. These populations are automatically analyzed by SAP2000 in parellel computing way. This parallel computing uses the beowulf cluster computing method (Fig. 3). Next, the GA processes are continued until the stopping criteria is reached and finally the result is found.

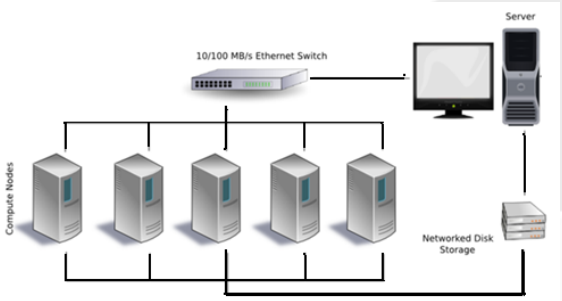


Figure 3. Beowulf concept for SAP2000-GA method to run in parallel computing [10].

3.1 Analyzed Steel Structure

There are three models which will be optimized in this research. The first model is 2D 10-bar truss optimization problem [7]. This model are also been analyzed by Duan [7], Cai and Thiereut [5], Coello [6], Rajeev and Krishnamoorthy [12] (see Figure 4).

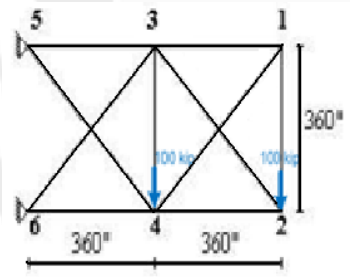


Figure 4. The first model 10 elements truss [7].

The second model of steel structure was 25 bar truss with specified dead loads [7]. This type also researces with different methods like Duan [7], Cai and Thiereut [5], Coello [6], Rajeev and Krishnamoorthy [12](see Figure 5).

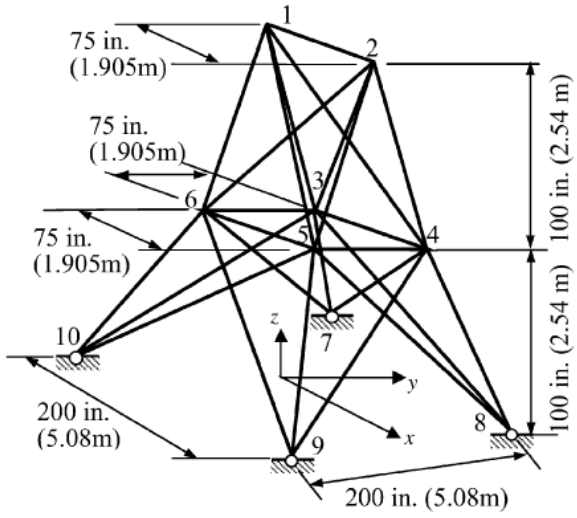


Figure 5. The second model 25 elements truss [7].

The third model (see Figure 6) is a 36-story irregular moment-resisting steel space frame structure with cross-bracings and an aspect ratio of 4.7.

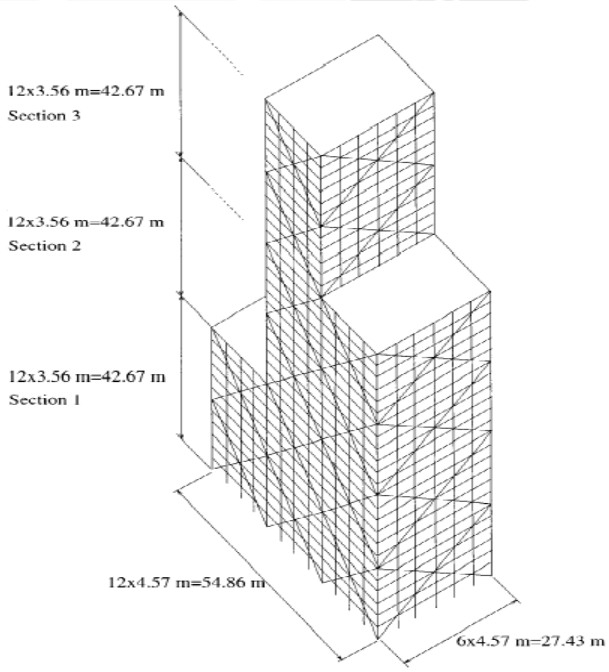


Figure 6. The third model 36-story steel frame structure [1].

A minimum weight solution for the same example was presented by [1] and [2]. It has 1384 nodes and 3228 members. The structure consists of three 12-story sections. In the lower sections 1 and 2, there are four groups of columns (Figure 6): corner columns, outer columns, inner columns in the unbraced frames, and inner columns in the braced frames. In section 3, only the first three types of columns are used. The beams in every floor are

grouped separately. In sections 1 and 2 they are divided into three groups: outer beams, inner beams in braced frames, and inner beams in unbraced frames.

The beams in section 3 are divided into two groups: inner and outer beams. In every three stories two different bracings are used with the same cross-section, one in the longitudinal direction and the other in the transverse direction. The inter-story drift is not limited. The dead load and the live load intensities on each floor are 2.88 kPa (60 psf) and 2.38 kPa (50 psf), respectively. The lateral loads due to wind are calculated with a basic wind speed of 113 km/h (70 mph), exposure C, and an importance factor of 1[1].

3.2 Define Objective Function

Main objective optimization of the first and the second model steel structure is minimizing structure's weight subjected dead load with constrained in member's stress and node displacement. After that we can create fitness function as describe below.

$$f_{(x)} = \sum_{i}^{n} \rho A_i L_i + C1 \sum_{i}^{n} g_{1i} + C2 \sum_{i}^{n} g_{2i} \qquad (4)$$

where ρ is material's density, A_i is member's cross section Area, L_i is member's length, C_1&C_2 are coefficient of constraint, g_{1i} and g_{2i} are penalty function due to following set of constraints:

if allowable stress > actual stress, then g1i = 0, otherwise g1i = 1. If allowable node displacement > actual displacement, then g2i = 0, otherwise g2i = 1.

For the third case, the objective optimization is to minimize structure's weight subjected specified load with constrained in member's ratio. The fitness function for the third case is

$$f_{(x)} = \sum_{i}^{n} \rho A_i L_i + C1 \sum_{i}^{n} R_{1i} \qquad (5)$$

where ρ is material's density, Ai is member's cross section Area, Li is member's length, C_1 is Coeficient of constraint, R1i is penalty function due to following set of constraints:

if allowable ratio > actual ratio, then $R_{1i} = 0$,
otherwise $R_{1i} = 1$.

Analysis process of SAP2000 results frame forces and node displacements. Design results data include the design stresses, stress ratios, effective lengths, optimal sections, and all other calculated quantities resulting from the design process. Three type data above are included in tho output file of SAP2000. Data will be processed to produce fitness value as part of GA procedure.

Before starting optimization process, coeficients used in eq. (4) and (5) must be defined experimentally.

4. ANALYSIS AND RESULTS

Two programs are created to solve the optimization problem. The first program is to optimize the 2D 10 elements and 3D 25 elements model, and constrained on the member stress and node displacements. The second program is to optimize the 3D 36 stories steel structures, where the design criteria is based on AISC-LRFD99 and node displacements. Figure 6 shows the flowchart of those programs.

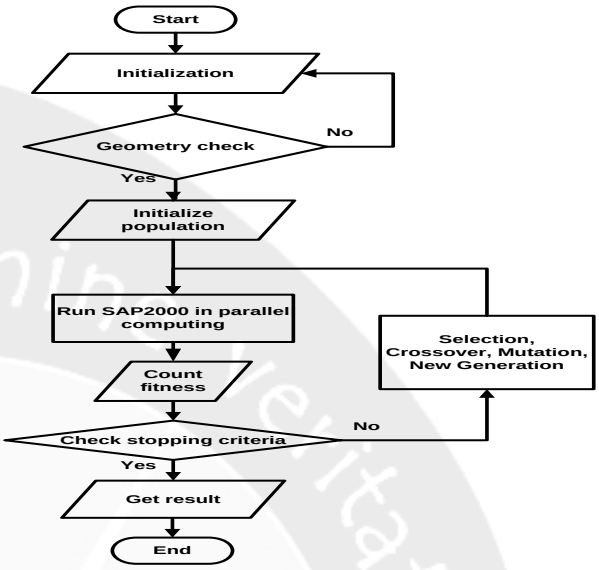


Figure 7. Flowchart for steel structure optimization.

After experimental study, it is found that the coeficient of objective function for the first and the second model is:

$$f_{(x)} = \sum_{i}^{n} \rho A_i L_i + 1000 \sum_{i}^{n} g_{1i} + 10000 \sum_{i}^{n} g_{2i} \qquad (6)$$

And for the third model:

$$f_{(x)} = \sum_{i}^{n} \rho A_i L_i + 100000 \sum_{i}^{n} R_{1i} \qquad (7)$$

The next step is to run the program with GA parameters showed in Table 1.

Table 1. Parameters used in SAP2K-GA method

	The first Model	The second Model	The Third Model
Population	100	100	200
Generation	200	200	1000
Cross over	0,8	0,8	0,8
Mutation	0,07	0,07	0,07
Member's constraint	25 Ksi	40 Ksi	AISC-LRFD99
Node's constraint	2 inches	0,35 inches	NA
Available options	32 binary cross sectional options	33 floating point cross sectional options	269 cross sectional options in 18groups

All optimization are executed in parallel computing method by using Beowulf cluster configuration (see figure 3) with one PC for running GA process and 10 PCs for running SAP2000.

Table 2. Comparison of (SAP2K-GA) with other techniques for the first truss model.

	Method				
	Optdyn	**Conmin**	**Coello**	**Rajeev**	**Sap2k-GA**
Weight (lbs)	5472-1	5563-2	5586-4	5613-5	5584-3
A1 (inch^2)	25.70	25.20	NA	33.5	33.5
A2(inch^2)	0.10	1.89	NA	1.00	1.00
A3(inch^2)	25.11	24.80	NA	22.00	23.20
A4(inch^2)	19.93	15.80	NA	15.50	18.20
A5(inch^2)	0.10	0.10	NA	1.620	1.00
A6(inch^2)	0.10	1.75	NA	1.620	1.00
A7(inch^2)	15.40	16.76	NA	14.20	18.20
A8(inch^2)	20.32	19.73	NA	19.90	21.39
A9(inch^2)	20.74	20.98	NA	19.90	21.50
A10(inch^2)	1.14	2.51	NA	2.60	2.20

Optimization results of the first model are compared with other techniques without any violation on member's stress and node displacements specified. This could be seen that the result of SAP2000-GA method (5584 lbs weight) is at the third place if compared with the other methods (See Table 2).

The optimization result for the second model shows that SAP2000-GA is at the third place with the total weight of structure 533,45 lbs and maximum deformation 0.15 inch (see Table 3).

Table 3. Comparison of (SAP2K-GA) with other techniques for the second truss model.

	METHOD				
	Chai &Thierut	**Rajeev**	**Duan**	**Der Shin**	**SAP2K-GA**
Weight (lbs)	487.28	545.86	562.78	485.17	533.45
A1(inch^2)	0.10	0.10	0.10	0.10	0.10
A2(inch^2)	0.10	1.80	1.80	0.30	0.80
A3(inch^2)	3.40	2.30	2.60	3.40	3.00
A4(inch^2)	0.2	0.10	0.10	0.10	0.10
A5(inch^2)	2.00	0.10	0.10	2.40	0.90
A6(inch^2)	1.00	0.80	0.80	1.00	0.90
A7(inch^2)	0.70	1.80	2.10	0.30	0.80
A8(inch^2)	3.40	3.00	2.60	3.40	0.34
Displ. max(inch)	0.14	0.14	0.14	0.15	0.15

The optimization result for the third model shows that SAP2000-GA is at the second place with the total weight of structure 20922.8 KN (see Table 4).

Table 4. Optimum structure weight obtained based on AISC-LRFD99 Code for the third model.

No	Type of Method	Weight of Structure
1	Adeli&Sarman(2006)	15410.1 KN - 15938.1 KN
2	Adeli&Park (1997)	21513.2 KN
3	SAP2k-GA	20922.8 KN

5. CONCLUSION

Two programs to optimize the steel structures have been created by combining SAP2000 and evolutionary algorithm by parallel computing The first model and the second model are analyzed based on the member stress and node displacement criteria and the third model is analyzed based on the AISC LRFD99. The result of this method is compared with other studies and shows that sap2000 successfully combined with GA

6. ACKNOWLEDGMENTS

Our thanks to Structure Engineering Laboratory of ITS Surabaya for allowing us to use computers for running these programs.

7. REFERENCES

[1] Adeli,H. and Sarma, K. C. 2006 Cost Optimization of Structures : Fuzzy Logic, Genetic Algorithms, and Parallel Computing, 2006 John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England.

[2] Adeli, H. and Park, H.S. 1997 Distributed Neural Dynamics Algorithms for Optimization of Large Steel Structures, Journal of Structurel Engineering, July 1997

[3] Computer and structures, Inc. 2000 SAP2000 - Static and Dynamic Finite Element Analysis of Structures. Berkeley. California, USA.

[4] Computer and structures, Inc. 2000 SAP2000 Integrated Finite Element Analysis and Design Structures : STEEL DESIGN MANUAL. version 7.4. revision May 2000. Berkeley, California, USA.

[5] Cai, J. B., and Thiereut, G., 1993 Discrete Optimization of Structures Using an Improved Penalty Function Method, Engineering Optimization, Vol. 21, No. 4, pp.293-306.
[6] Coello, C.A.C. Discrete Optimization of Trusses using Genetic Algorithms, http:\\delta.cs.cinvestav.mx/~ccoello/conferences/expersys94.pdf.gz
[7] D.S. Juang et al. 2003 Optimum Design of Truss Structures Using Discrete Lagrangian Method, Journal of the Chinese Institute of Engineers, Vol. 26, No. 5.
[8] Gen, M. and Cheng, R. 1997 Evolutionary Algorithm and Enginering design. A wiley-Interscience publication, John wiley & Sons, Inc., New York.
[9] Goldberg,D.E .1989. Evolutionary Algorithm in Search,Optimization and Machine Learning. Addition wesley publishing company,Inc,USA.
[10] Haupt, R. L. And Haupt, S.E. 2004 Practical Genetic Algorithms, 2nd Edition, A Wiley Interscience publication
[11] Khennane, Amar . 2005 Performance Design Of Reinforced Concrete Slabs Using Commercial Finite Element Software . http://eprints.usq.edu.au/708/1/Khennane_SLAB_DESIGN_revised_paper.pdf
[12] Rajeev, S. and Krishnamoorthy, C. S .1992 Discrete optimization of structures using evolutionary algorithms", in Journalof Structural Engineering 118(5), 1992, pp. 1233-50.
[13] Setareh, M., Bowman, D. A.,and Tumati, P. .2001 Development of a collaborative design tool for structural analysis in an immersive virtual environment. Seventh International IBPSA Conference, Rio de Janeiro, Brazil, August 13-15, 2001
[14] Sesok, D. and Belevicius, R. 2007 Use of GA in Topology Optimization of truss Structures. ISSN 1392 – 1207, MECHANIKA 2007. Nr.2(64).

The Hydrophobic-Polar Model Approach to Protein Structure Prediction

Tigor Nauli
Research Center for Informatics, Indonesian Institute of Sciences (LIPI)
Jalan Cisitu, Sangkuriang
Bandung 40135, INDONESIA
Tel. +62 22 2504711
tigor.nauli@lipi.go.id

ABSTRACT
Protein folds into a specific native three-dimensional structure to form its functionality. The prediction of a protein structure from its amino acid sequence is one of the most important problems in computational biology. One abstraction of the problem is so called Hydrophobic-Polar (HP) model, which is searching the maximum number of non-consecutive pairs of hydrophobic amino acid that stable conformation of low free energy.

A Genetic Algorithms search procedure was developed for use in protein folding prediction. During the generation steps, a population of conformation of the protein is maintained, conformation are changed by mutation, and crossover in some parts of amino acid sequence are interchanged between conformation. The conformation of protein is represented by a sequence of moves on a cubic grid uses the HP-model.

Employing the GA-based technique to predict the structures of the short proteins yield the optimal conformations with lower energy minimum than previously reported. Further improvement to the HP-model should include other characteristics of amino acids, as the realistic structure of the protein is not necessarily the optimal structure predicted.

Keywords
Genetic algorithm, HP-Model, protein structure prediction, protein folding.

1. INTRODUCTION
A protein is a chain of amino acid residues that folds into a specific native three-dimensional structure under natural conditions. Functionality of a protein is mainly defined by its 3D fold. Protein folding is driven by a diversity of forces, including covalent, van der Walls, and hydrogen bonding. The variety and complexity of protein's folds requires more advanced methods in the predicting the protein structure [5].

Currently, protein structures are primarily determined by techniques such as NMRI (nuclear-magnetic resonance imaging) and X-ray crystallography, which are expensive in term of equipment, computation and time. They also require isolation, purification and crystallization of the target protein. Therefore, computational approaches to protein structure prediction are very attractive.

One computational approach in predicting protein structure is based on contacts of an amino acid sequence. Amino acids are divided into hydrophobic (H) and polar (P) ones. In the process of forming a tertiary structure, the hydrophobic amino acids tend to form non-covalent bonds since these leads to stable conformation of low free energy. The HP-model is a lattice model. The amino acids of a protein are placed on the vertices of a grid such that consecutive amino acids are placed on the grid side by side. In this approach, a fold of the protein into a square or cubic grid is searched that exhibits the maximum numbers of non-consecutive pairs of hydrophobic amino acids in direct contact [7].

The protein folding problem is known to be NP-Complete in both two-dimensional and three-dimensional square lattices [1]. It has been shown that protein folding, at least on a lattice, is a member of the class of NP-complete problems. Therefore there is probability exists no general search algorithm that can be guaranteed to find the global free energy minimum for real proteins [8].

A number of well-known heuristic optimization has been applied to the two-dimensional HP Protein Folding problem, including Evolutionary Algorithms [8], Monte Carlo (MC) algorithm [5], and Ant Colony Optimization algorithm [6]. A simple approach using Genetic Algorithm was also introduced to optimize protein structure [2]. It included some protein characteristics in more-complex models to predict the three-dimensional chains of residue.

In this work, we implementing the Hydrophobic-Polar model in a three-dimensional cubic lattice using a Genetic Algorithm (GA) as a tool to find the optimal conformation for given the amino acid sequence. The residue location on the grid is represented by absolute encoding of move sequences. The results are compared with real short-length proteins with known folding to justify the prediction.

2. METHODS
In implementing the GA for optimizing protein structure prediction, we prepared a protein representation scheme, designed a fitness measurement method, and chosen termination criteria. The computational effort was help by a GA program written in Java to conduct the prediction.

2.1 Protein Representation
The protein was usually characterized by a string of letters, which denote the amino acid residues. Based on its hydrophobicity, the string was translated into an array of letter 'H' (for hydrophobic residue) and letter 'P' (for polar residue) according to Table 1.

Non-standard amino acid symbols, such as X and Z, were skipped in this translation.

Table 1. Hydrophobicity of amino acid residues.

Hydrophobic Amino Acid	
Alanine	A
Glycine	G
Isoleucine	I
Leucine	L
Methionine	M
Phenylalanine	F
Proline	P
Tryptophan	W
Valine	V

Polar Amino Acid	
Arginine	R
Asparagine	N
Aspartic acid	D
Cysteine	C
Glutamic acid	E
Glutamine	Q
Histidine	H
Lysine	K
Serine	S
Threonine	T
Tyrosine	Y

We used a string of *moves* on the grid, to represent a protein conformation on the lattice. The moves being represented in an obvious manner as directions 'U' (up), 'D' (down), 'L' (left), and 'R' (right). It was also defined directions 'B' (backward) and 'F' (forward) to represent move away or toward the last step before, as illustrated in Figure 1.

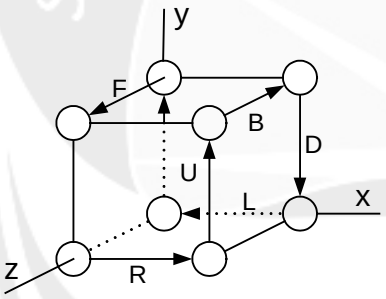


Figure 1. Definition of moves.

The binary representation for a protein is L*4 bits long, where L is the number of amino acid residues in the protein. A group of 4 bits is decoded to an integer between 0 and 15 as genes. Some moves are redundant. Each move along the cardinal direction differs by one integer.

2.2 Experimental Parameters

Four GA parameters: initial population size, number of generations run, crossover rate, and mutation rate, are pre-determined before the prediction to achieve best results. The initial population size was set to 1000. The optimal values were found to be 0.7 for the crossover rate and 0.001 for the mutation rate. The crossover operation is repeated until (N – 1) new accepted structures have been constructed to constitute the population of the next generation. The GA was run for 200 generations

The 20-residue long sequence HPHPPHHPHPPHPHHPPHPH is selected to benchmark the GA. The optimal energy for the sequence is -9, as reported by Unger and Moult [8].

We chosen a short-length protein, Kappa-Hefutoxins, a toxin found in scorpion venom, as test protein. Its known structure has been determined experimentally. The characteristic of the protein was downloaded from Protein Data Bank as structure file 1HP9 [4]. The protein sequence is GHACYRNCWREGNDEETCKERC.

2.3 Termination Condition

In every step of generation the chromosome, we measure the fitness of a given structure. All non-viable protein structures, such as more than one residue laid on the same coordinate location, are eliminated from the chromosome. The algorithm terminates when the free energy evolved was reach the expected target, no other lower energy found in a period of time, or the generation step exceeds the CPU cut-off.

2.4 Visual Analysis

In order to analyze the results of the optimization, we developed a simple protein structure viewer in Java. This viewer consists of a Java program to translate the binary representation of a protein structure into a set of coordinates in three-dimensional integer space and a Java 2D API Graphics to graph these coordinates and the chain connecting them.

3. RESULTS AND DISCUSSION

In predicting the 20-residue long sequence, our GA found four structures with energy of -9 and one structure with lower energy -10. The folding of the later structure is represented in Figure 2. All the hydrophobic residues tend to be inside of a low energy structure, while the hydrophilic (polar) residues are forced to the surface. Low energy conformations are compact structures maintaining a hydrophobic core. The optimal conformation with energy -10 was found in this 3D HP-model. Contrarily, for the same 20-residue, the 2D HP-model protein structure prediction yielded the optimal conformation with lowest energy of –9, as discovered by some authors [2, 8].

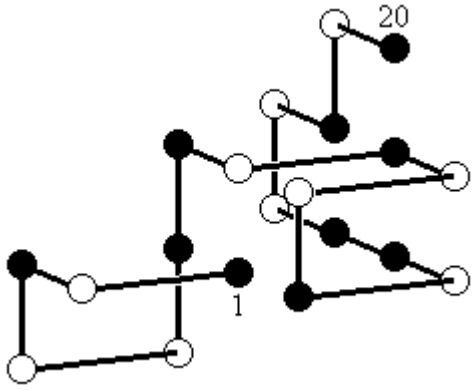


Figure 2. Conformation of a protein with energy of -10.

The Hefutoxins composed of four hydrophobic residues. It can be expected that the optimal structure of the protein have the lowest energy -3. The GA prediction found the structure of Hefutoxins with the optimal folding conformation as in Figure 3. This conformation shows a small protein molecule with a hydrophobic head and a long hydrophilic tail. The small amount of hydrophobic residues in Hefutoxins formed a planar protein rather than globular shaped as natively.

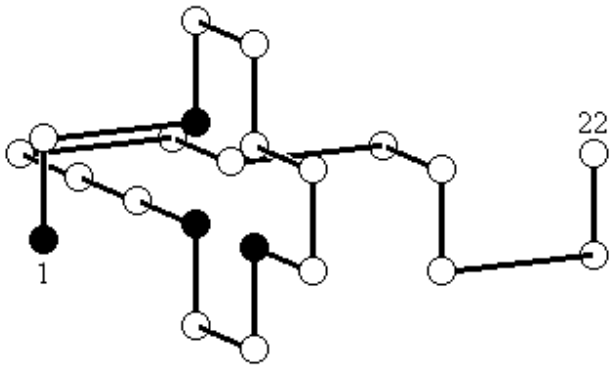
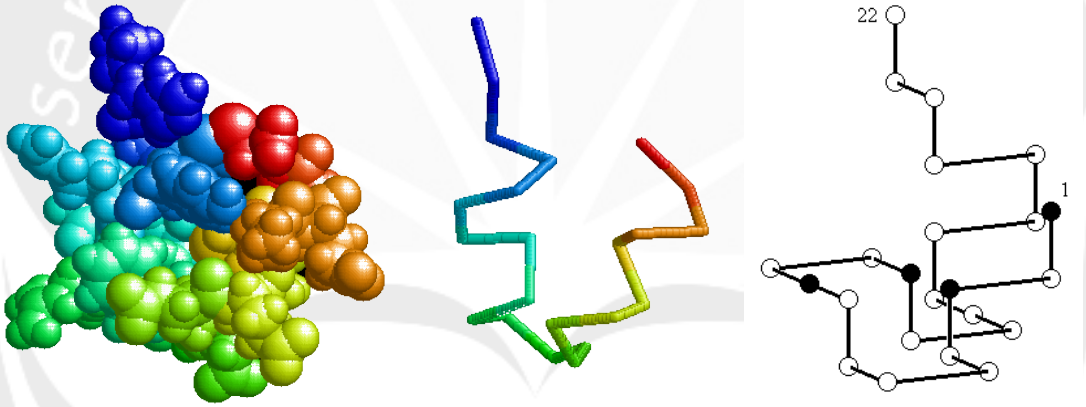


Figure 3. The optimal structure of Hefutoxins with energy of -3.

We observed that the predicted protein structure by the GA different than the real folding structure. In fact, the Hefutoxins will fold into a helical conformation in two separate amino acid chains. Such structure emerges early in the process, that is in the chromosomes' generation before the optimal structure is reached. A structure with energy of -2 shows the closest shape as the real Hefutoxins' conformation, as seen in Figure 4.



(a) (b) (c)

Figure 4. The tertiary structure of Hefutoxins (a), its backbone structure (b), and the predicted protein structure by the GA (c).

Two pairs of non-consecutive hydrophobic residues are bind and leaves the other residues arranged in two helical chains.

Comparing with the conformation of the 20-residu long protein, which is globular and has a global energy minimum, the structure of Hefutoxins exhibited an example of local energy minimum. It is not known whether the functional conformation of a globular soluble protein is necessarily at the global free energy minimum [3].

The main disadvantage of heuristic methods, as it is mentioned by some authors, is that they achieve good folding for short proteins only. Our GA is not implemented yet to predict the structure of longer protein, however we realized that the other characteristics of protein could influence the HP-model. Some of the characteristics are molecular size that hinders one residue being place close to other residues, and opposite charges between two residues that causing a fold. It is encouraging, without making major revision to our GA, to expand the HP-model to be implemented into a triangular grid, to achieve more realistic folding.

4. CONCLUSION

The three-dimensional HP-model approach to short protein prediction, employing Genetic Algorithms, improves the method to find the optimal conformation and predicts protein structure with lower energy minimum. The real conformation of short protein is not necessarily at the optimal structure predicted. We should include the other characteristics of protein in the HP-model and uses other form of grid to improve the folding algorithm.

5. REFERENCES

[1] Berger, B and Leighton, T. 1998. Protein folding in the hydrophobic-hydrophilic (HP) model is NP-complete. In Proceedings of the Second Annual International Conference on Computational Molecular Biology (USA, March 1998).

RECOMB '98. ACM Press, New York, NY, 30–39. DOI= http://doi.acm.org/10.1145/279069.279080

[2] Braden, K. 2004. A simple approach to protein structure prediction using genetic algorithms. Course Notes. Standford University Press.

[3] Levinthal, C. 1968. Are there pathways for protein folding? J. Chem. Phys. 65, 44-45.

[4] PDB. 2009. Protein Data Bank, http://www.rcsh.org/pdb (Last accessed 21 November 2009)

[5] Sali, A., Shakhnovich, E. and Karplus, M. 1994. How does a protein fold? Nature 369, 248-251.

[6] Shmygelska, A. and Hoos, H. H. 2005. An ant colony optimisation algorithm for the 2D and 3D hydrophobic polar protein folding problem. BMC Bioinformatics 6:30.

[7] Sperschneider, V. 2008. Bioinformatics: Problem Solving Paradigms. Springer-Verlag.

[8] Unger, R. and Moult, J. 1993. Genetic algorithms for protein folding simulation. J. of Mol. Biol. 231, 75-81.

University Course Scheduling Using The Evolutionary Algorithm

Ade Jamal
Informatics Engineering, University Al-Azhar Indonesia
Jl. Sisingamangaraja 1, Kebayoran baru, Jakarta 12110
(62-21)7244456
adja@uai.ac.id

ABSTRACT
Course scheduling problem is hard and time-consuming to solve which is commonly faced by academic administrator at least two times every year. This problem can be solved using search and optimization technique with many constraints. This problem has been well studied in the past, and still becomes favorite subject for researchers. We will briefly discuss the convergence difficulty in our initial work on this subject using a modified hill-climbing search technique[8]. In this paper, an evolutionary algorithm is applied to solve the course scheduling problem and studying mutation techniques involved in the algorithm.

Keywords
Course Scheduling, Optimization, Evolutionary Algorithms Genetic Algorithms

1. INTRODUCTION
Building a course schedule is one of the main challenge at university that must be faced by academic administrator every semester. This problem is considered as an NP-complete problem [2,3] which is quite difficult and time-consuming to solve. This problem has been the subject of extensive research effort due to its complexity and wide application such as school timetables [1], exam scheduling [3] and course scheduling [2,5,8,9,11]. The importances of seeking for good technique to solve this kind of problem are realized by some educational institutions by organizing International Timetabling Competition[1].

The university course scheduling problem is the task of assigning the academic events (such as lectures, tutorials etc) to room and time slots in such a way that taking consideration a predefined set of constraints. Every university may have different constraints. However these constraints usually can be classified as two types, namely hard and soft constraints. Hard constraint must not be violated to construct a valid or feasible schedule, while soft constraints are desired but not absolutely to be fulfilled.

A number of algorithms have been used to solve the course scheduling problem. The graph coloring heuristic techniques whereby course are assigned to rooms and time-slots one by one in particular order are the earliest approaches which are very efficient in small scheduling problem. The second approaches are the local search algorithm family that basically perform search in neighborhood of a known solution state rather than exploring possible solution in wider search space. The most popular local search method are simulated annealing method [1] and tabu search method [5]. The third type is the evolutionary algorithms and genetic algorithms which is based on Darwinian evolutionary theory [3,9,11].

In the current work we use evolutionary algorithm to solve the course scheduling problem, focusing more deeply in reproduction mechanism. This paper is organized as follows: section two described the course scheduling problem based on constraints and its representation model. Section three discuses the algorithms and its applicability to the course scheduling problem. Section four discusses the experimental results and compares it to the previous work [8]. In the last section we present a conclusion and recommendation for future work.

2. COURSE SCHEDULING PROBLEM
2.1 Problem definition
Despite of different constraints, definition of university course scheduling problem can vary depend on whether it is based on post student enrollment where a set of student attending each event are defined [2,9,11] or based on curriculum for each faculty where scheduling takes place prior enrollment [10,11]. Since our university (i.e. University Al-Azhar Indonesia) conforms to the curriculum based scheduling, our work stick with it.

Hence the university course scheduling problem is defined as follows: There is a set of room which has a seat capacity and contains specific feature (i.e. laboratories), a set of course which is based on curriculum for each program and may have multiple section (i.e. credit unit where one course section occupy one time-slot), a set of lecturer which has been assigned to specific course(s) and has a certain unavailable time-slot. A set of events (i.e. classes), to be scheduled in a certain number of time-slots or time-period and a room, is defined as combination of a specific course with assigned lecturer attended by certain number of student from a particular group (i.e. a student group come from a specific program and same grade).

A feasible schedule is one in which all the classes have been assigned to a time-slot and a room whereby the following hard constraints are satisfied:

1. rooms must not be double booked for classes at any feasible time-slot (classroom clashed)

[1] The International Timetabling Competition 2002 - http://www.idsia.ch..
The International Timetabling Competition 2008 sponsored by PATAT and WATT (http://www.cs.qub.ac.uk/itc2007)

2. the room capacity must not be exceeded by the number of attending student
3. the room has a feature (i.e. laboratory) required by the classification
4. a lecturer can not teach more than one one class at the same time
5. a lecturer can not teach any class in time-slot which is unavailable for him/her
6. a student group from the same program and same grade can not attend more than one class at the same time
7. a class with multiple section must be assigned in the same room contiguously.

A number of soft constraints that should be minimized could introduced such as a lecturer should be assigned in his/her preference time-slot and room or parallel classes (i.e. the same courses taught by different lecturer) should be scheduled at the same time. However, in the present work we will not take into account soft constraints. The main objective is first to find a feasible course schedule, and then this feasible schedule to be optimized with respect to soft constraints in the next phase.

2.2 Scheduling Model

Deducted from previous published works, there are also two different approaches were used to represent the scheduling model. In the first approach, the scheduling model is represented by a two dimensional matrix where each row corresponds to room, each column to a time-slot, and then the matrix element contains a particular event or blank. This approach is usually used for post-enrollment scheduling problems [2,3,9,11]. In the second approach, a scheduling model is represented directly by a triple of <event, room, time-slot>. Note that an event in the second approach is a combination of course and initially assigned teacher. The second approach is found in the prior-enrollment scheduling publications [10,11].

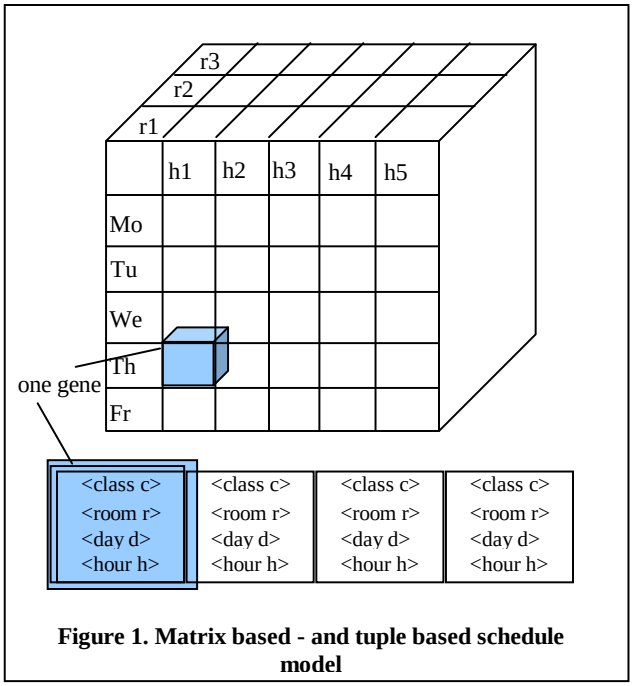


Figure 1. Matrix based - and tuple based schedule model

Though these two different approaches corresponds respectively to two different scheduling problems, but none of the previous authors explained the correlation between the model and the problem, even Pawel Myszkowski [11] who considered both problem variations and used both approached respectively,

The presented work invoke a quite different approach. Both kind of approaches (i.e. matrix and an array of tuple) are used for the sake of ease evolutionary mutation and fitness evaluation. Furthermore instead of using a two-dimensional matrix we explode it to three-dimensional matrix by remodeling a one-dimensional column which represents time-slot into a two-dimensional matrix where each row represent day and each column as hour. In contrary to some previously published work [3,9], where the length of time slot can vary, here the size of matrix is predefined by the problem. Hence, the first hard constraint is completely satisfied (i.e. only one class is scheduled in each at any feasible time-slot).

3. EVOLUTIONARY ALGORITHM

Evolutionary algorithms are population based meta-heuristic optimization algorithm inspired by biological evolution that used mechanisms like mutation, crossover, natural selection and survival of the fittest in order to refine a set of solution candidates iteratively. The first version of the evolutionary algorithm introduced by Rechenberg (1965) where he began with a parent and a mutated one, whichever is the fittest became a new parent [7], hence only one species involves in every generation. In our previous work, we used the same strategy with exception that the number of mutant version is larger [8]. Although the convergence rate to the solution is found to be very fast in the previous work, but this kind of local search is often stuck in a local optimum. In the current work, we will incorporate population for each generation in order to get the necessary diversity in the solution candidates.

The matrix based and array of tuple based scheduling models defined in the previous section represent an individual chromosome where an element of matrix or a single tuple represents a gene as shown in figure 1. A kind of permutation encoding is used for this purpose where each matrix element or each tuple representing a gene will contain an event or class index. Noted that a course taught in more than an hour will be presented in multiple section or multiple genes of same indexes. The permutation encoding in matrix form reveals a straight forward decoding, or strictly speaking no decoding is necessary.

Using a permutation like encoding has another issue in the crossover mechanism, namely, we have to check whether genes in the half chromosome from one parent are also found in the other half chromosome from the other parent. This extra work requires a large computation effort. Therefore, a crossover-like mechanism within one chromosome is applied, namely swapping a sub-set of scheduled classes. However since this mechanism involves only one parent, precisely speaking it is a mutation mechanism.

Initial population is generated by randomly mutating one very first chromosome as many as number of population. The first chromosome is constructed in such way that all events are scheduled in the schedule matrix and take into account that classes with multiple section are scheduled on the same room and

contiguously. Hence it fulfill the first and the seventh hard constraint in the previous section. These two hard constraints are not only used in constructing the first chromosome but also considered in the mutation mechanism, thus we name them fixed constraints. The algorithm used for constructing the first chromosome is based on greedy algorithm by course hours.

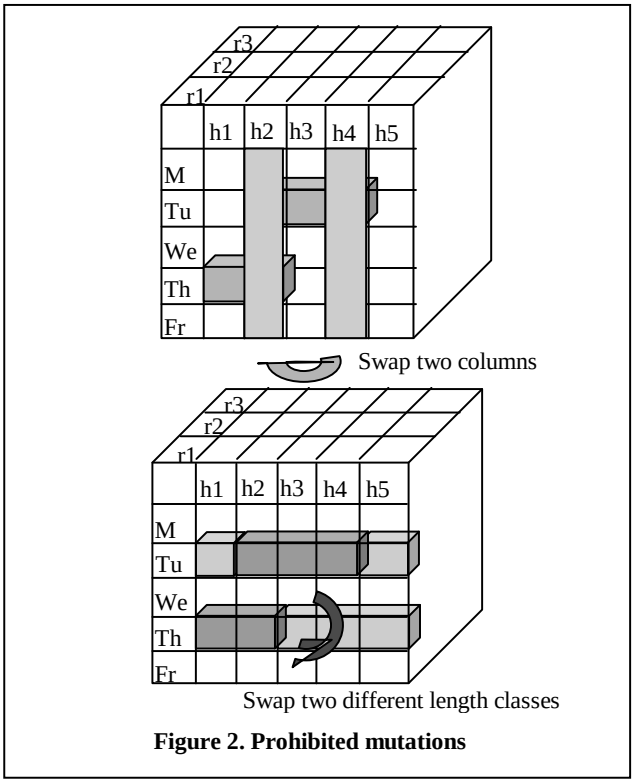


Figure 2. Prohibited mutations

Consequences of considering fixed constraints in the mutation scheme, swapping column wise (changing hour) is prohibited because it would destroy the contiguity of multiple section (multiple hour) classes. Swapping two classes with different number of sections (hours) is sometimes impossible in case insufficient time-slot for longer class in the new room and day to be placed. These two prohibited mutations are shown in figure 2. As in our previous work [8], only two mutation operators are used, namely swapping two arbitrary events (classes) and swapping all events that scheduled on two arbitrary days and rooms.

The fitness of each chromosome depends on the amount of violated hard constraints. A value of 100 is added to the fitness if all event completely comply to one type of hard constraint. This maximum fitness value decreases proportionally with the amount of constraint violation. Every constraint will be evaluate separately for its fitness. Since there are 5 hard constraints left in our problem, the maximum total fitness value to be sought is 500.

All chromosome in the population are ranked according to its fitness. A group of elite chromosome, i.e. the most fittest will be kept for the next generation. The number of this elite group is only a small fraction of the population size. The other survival will come from further evolutionary mechanism, namely selection of parents based on roulette wheel techniques, and then applying both mutating operators with certain probabilities on the selected parents. Replacement of the least fittest chromosomes by elite group finalizes a creation of new generation, and this process is repeated until maximum fitness (valid solution) is found.

4. EXPERIMENTAL ANALYSIS

For experimental analysis purpose we take a relative small test set of 75 classes (179 course hours) to be scheduled in 4 rooms (192 hour available time-slots). Using local search algorithms (i.e. modified hill-climbing search) from our previous work [8], this test set yields about 40% convergence failure. However, it need only less than 50 iterations or generation when it successfully converged as depicted in figure 3. The reason of high rate of convergence failure is lack of diversity and relatively narrow exploring area search in the used local search techniques which causes the process reaches local maximum quickly.

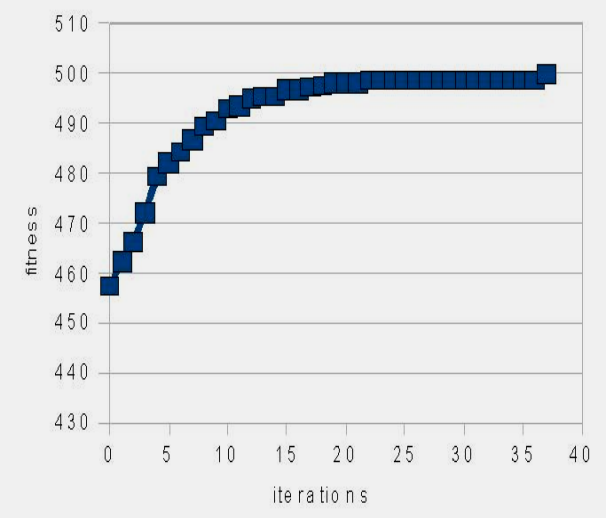


Figure 3. Fitness variation as function of generations produced by modified Hill-climbing search [8]

Incorporating a group of individuals or a population in the present work induces some parameters to be determined. Those parameters are number of population Np, probabilistic rates for mutation of swapping two classes Pc, and probabilistic rates for mutation of swapping two scheduled days/rooms Pd, and percentage of elitism Pe. The first three parameters influence the diversity force of the evolutionary algorithms while the percentage of elitism and together with selection techniques effect the force of pushing quality [4]. Effect of population size is very clear, namely larger size yields more diversity. Percentage of elitism is more or less also quite predictable, namely too much elitism causes less diversity. In this study we took Np=200 and relative small elitism Pe=20%.

The other two parameters, regarding mutations probabilistic, which can result in good convergence rates should be further studied to be chosen.

Hence the objective of present experiment are studying mutation parameters. Fixing the probabilistic rate of swapping two scheduled days/rooms Pd=60%, effect of the probabilistic rate of swapping two classes Pc is studied. We run the program a couple times until either the maximum fitness was found or the limit number of generation was reached for a certain value of PC, and then repeat the process by varying the Pc. The following table presents the statistic results acquired from twelve runs for each parameters value of Pc. The values shown in the table are the number of generation needed before maximum fitness value was reached for four basic statistic parameters (I.e average, standard deviation, min and max) and the percentage of failure attempts from the first twelve runs.

Table 1. Effect of probabilistic rate of swapping two classes on number of generations

Pc	1	0.8	0.6	0.4	0.2
Average	295	438	645	574	710
Std Dev	191	469	498	301	292
Min	156	179	240	270	352
Max	871	1771	1620	1318	1176
Failure	0	0	0	0	50.00%

Fixing the probabilistic rate of swapping two classes Pc=60% we vary the Pd as shown in the following table

Table 2. Effect of probabilistic rate of swapping two scheduled day on number of generations

Pd	0.6	0.4	0.2	0.1	0
Average	645	267	186	344	551
Std Dev	498	138	37	312	355
Min	240	144	135	144	245
Max	1620	630	256	1081	1190
Failure	0	0	0	0	30.00%

The results has shown that deviation and average generation significantly effected by these two parameters Pd and Pc. Mutation by swapping two classes (Pc) is required more than mutation by swapping two scheduled day. Furthermore, Pc less than 20% results in convergence failure rate above 50% , while Pd =0 (no mutation by swapping two scheduled day) still yields successful convergence about 70% from twelve attempts. This happened because the swapping two classes is the diversity force for exploitation in detail of successor solution state, while swapping two scheduled day for exploration in wider solution space around the successor. Our test has shown the Pd=20% and Pc=60% give the best result so far.

Experimental results have also shown that the fitness value approaches quickly to the maximum value after about the average number of generations as depicted in figure 4. Thereafter the fitness curve flattens up until certain generation increases again to maximum values. This phenomenon is not found in another run for the same case (i.e. the same test set and same evolutionary parameters) where maximum fitness value is found at about average number of generations as pictured in figure 5. In this case, the fitness values is steadily increasing.

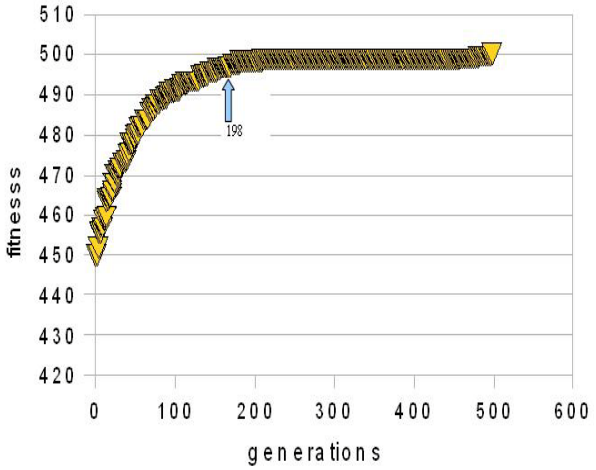


Figure 4. Fitness variation as function of generations produced by successful run after 498 generations, for Pc=1.0, Pd=0.2 where average number of generation is 198.

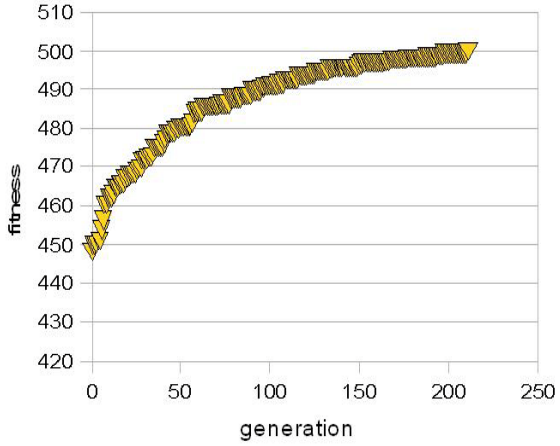


Figure 5. Fitness variation as function of generations produced by successful run after 211 generations, for Pc=1.0, Pd=0.2 where average number of generation is 198.

5. CONCLUSION

We have presented the evolutionary algorithm for solving of university course scheduling problem. This study has shown that the probabilistic rate of two mutation types has significant effect on successful rate of the algorithm. This evolutionary algorithm can find feasible or valid course schedules very good by using small probabilistic rates for mutation of swapping two scheduled

days/rooms Pd and large probabilistic rates for mutation of swapping two classes Pc.

The mutation by swapping two scheduled days/rooms which is the force for exploration in wider solution space around the successor can good replace the real crossover mechanisms between two parents.

The mutation by swapping two classes is absolutely needed to ensure convergence. This is the force for exploitation in the vicinity of global optimum solution. Some experimental results have shown that this exploitation force need more time in one case than the others.

For future work, we aim to take into account the soft constraints by introducing second stage local optimization.

6. REFERENCES

[1] Abramson, D. (1991) Constructing School Timetables using Simulated Annealing: Parallel and Sequential Solutions", Management Science, Vol. 37, No. 1, , January, 1991, 98-113

[2] Al-Betar M.A., Khader A.T. and Gani T.A. (2008) A Harmony Search Algorithm for University Course Timetabling. In: Burke E., Gendreau M. (eds.). The Proceedings of the 7th International Conference on the Practice and Theory of Automated Timetabling, Montréal, Canada, 2008.

[3] Burke, E.K., Elliman, D.G. and Weare, R.F. (1994) A Genetic Algorithm based University Timetabling System, In Proceedings of the 2nd East-West International Conference on Computer Technologies in Education, Sept, 1994, Crimea, Ukraine, 35-40

[4] Eiben, A.E. and Smith, J.E. (2007) Introduction to Evolutionary Computing, Natural Computing Series 2nd Edition, Springer

[5] Elloumi, A., Kamoun, H. and Ferland, J.(2008) A Tabu Search for Course Timetabling Problem at a Tunisian, in Proceeding of the 7th International Conference on the Practice and Theory of Automated Timetabling PATAT '08, Edmund K Burke and Michel Gendreau (eds), August 2008

[6] Elmohamed, M.A.S., Fox, G. and Coddington, P.(1998) A Comparison of Annealing techniques for Academic Course Scheduling", DHPC-045, SCSS-777, 1998

[7] Haupt, R.L. And Haupt, S.E.(2004) Practical Genetic Algorithms, Wiley-InterScience 2nd Edition

[8] Jamal, A. (2008) Solving University Course Scheduling Problem using Improved Hill Climbing Approach, In Proceeding of the International Joint Seminar in Engineering, August 2008, Jakarta, Indonesia

[9] Lewis, R. and Paechter, B. (2005) Application of the Grouping Genetic Algorithm to University Course Timetabling, In G. Raidl and J. Gottlieb (eds) Evolutionary Computation in Combinatorial Optimization, Berlin Germany, Springer LNCS 3448, pages 144-153

[10] Moody, D., Kendall, G. and Bar-Noy, A.(2008) Constructing initial neighborhoods to identify critical constraints, In Proceedings of the 7th International Conference on the Practice and Theory of Automated Timetabling PATAT '08, Edmund K Burke and Michel Gendreau (eds), August 2008

[11] Myszkowski, P. and Norbeciak, M. (2003) Evolutionary Algorithms for Timetable Problems. Annales UMCS Informatica AI, 2003, 115-125. DOI= http://www.annales.umc.lublin.pl/

Adaptive Appearance Learning Method Using Simulated Annealing

Du Yong Kim
Gwangju Institute of Science and Technolgy
261 Cheomdan gwagiro, Buk-gu Gwangju,
Mechatronics building
+82-62-715-3266, Republic of Korea
duyong@gist.ac.kr

Ehwa Yang
Gwangju Institute of Science and Technolgy
261 Cheomdan gwagiro, Buk-gu Gwangju,
Information & communication building
+82-62-715-2406, Republic of Korea
mgjeon@gist.ac.kr

Vladimir Shin
Gwangju Institute of Science and Technolgy
261 Cheomdan gwagiro, Buk-gu Gwangju,
Mechatronics building
+82-62-715-2397, Republic of Korea
vishin@gist.ac.kr

Moongu Jeon
Gwangju Institute of Science and Technolgy
261 Cheomdan gwagiro, Buk-gu Gwangju,
Information & communication building
+82-62-715-2406, Republic of Korea
mgjeon@gist.ac.kr

ABSTRACT
In this paper, we propose an adaptive appearance model for a visual tracking application. In appearance-base visual tracking, we solve the filter drift problem caused by the adaptation toward non-targets. Simulated annealing method is proposed to give the adaptive property in appearance subspace learning so that the filter drift problem is able to be effectively avoided. Experimental results from real-video sequence illustrate the robust performance of the proposed idea.

Keywords
Computer vision, visual tracking, particle filter, appearance subspace learning

1. INTRODUCTION
Problems associated with visual tracking have been heavily investigated in the computer vision community. Despite many algorithms being suggested from this body of work, challenging issues remain when they are implemented in real-world circumstances. These challenges mainly stem a primary source: the design issue of the robust observation function. Among the many types of tracking frameworks, we focus on the appearance-based tracking because it is widely used and more dependent on these issues than other approaches.

In the conventional particle filtering for visual tracking [1], we draw a set of particles from the predicted state density which is the distribution of predicted locations of the object. If theses predicted locations are not determined precisely, there will be distractions in the filter; the filter distractions will then gradually adapt to the non-targets, resulting in filter drift. The main source of the filter distraction is an incorrectly cropped target appearance. In such cases, we can intuitively overcome this problem by exploring a sufficient number of samples in dense area. However, this method requires heuristics and a high computational cost. In previous research, the optimal importance function [2], and the discriminative observation system design approaches [3, 4] have been discussed both independently and together to solve this problem.

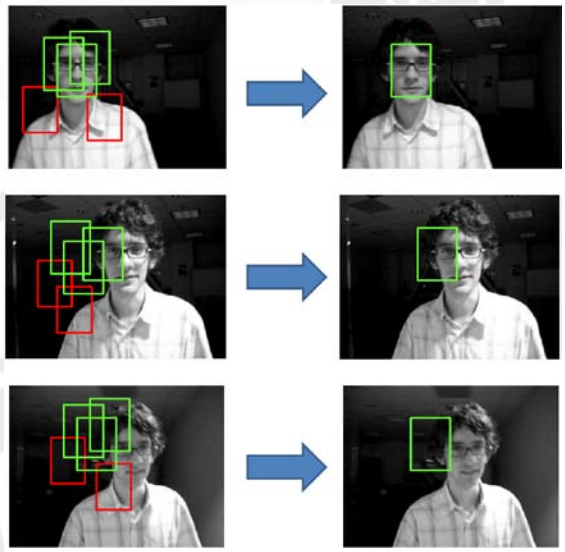


Figure 1. Appearance-based tracker failure due to the inaccurate observation evaluation

In appearance-based tracking, the template (reference appearance) update is recently proposed that learns principal components (eigenspaces) from a set of observations [5]. To reflect

history of previous observation, subspace learning method for appearance has been proposed based on incremental principle component analysis (IPCA). Under the IPCA, particle filter (PF) has shown robust performance against temporal occlusion, illumination change, and pose change. However, IPCA-based PF suffers from distraction problem when the target appearance has changed rapidly or a set of samples does not include the correct one. When the filter starts to have distraction, then mis-aligned appearance occurs which leads to filter drift. In Figure 1, the filter

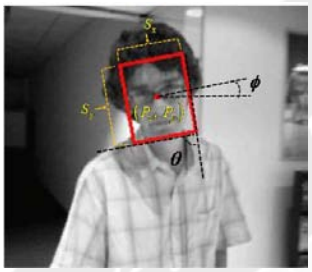


Figure 2. Local coordinate based appearance model

gradually drifts when a sample set does not have the best fit one.

To solve the problem, robust observation function is proposed. Adaptive property is incorporated in subspace update process by using simulated annealing. From simulated annealing, the accumulation error from mis-aligned appearance is reduced so that the distraction is alleviated in advance.

The remainder of paper is organized as follows. Section 2 describes the formulation of problem. Section 3 discusses the incremental visual tracking (IVT) [5] as a basic framework. Simulated annealing and proposed idea is explained in Section 4. Experimental results are provided in Section 5. Finally conclusion is made in Section 6.

2. STATEMENT OF THE PROBLEM

Let the state vector x_t represent components in the local coordinate based approach, in other words, x- and; y- position of the box center $p=(p_x, p_y)$, the scale of the box $S=(S_x, S_y)$, the rotation angle ϕ, and the skew direction θ of an object as described in Figure 1. Then, the aim of probabilistic tracking is to estimate x_t based on the probability density function of x_t given through the observation set $Z_t=\{z_1,...z_t\}$ and described as two Bayesian recursion equations.

Prediction: $p(x_{t+1}|Z_t)=\int p(x_{t+1}|x_t)\,p(x_t|Z_t)\,dx_t$

Update: $p(x_{t+1}|Z_{t+1}) \propto p(z_{t+1}|x_{t+1})\,p(x_{t+1}|Z_t).$ (1)

In the prediction step, a set of bounding boxes (sample) are chosen from the state transition density $p(x_{t+1}|x_t)$. The observation likelihood $p(z_{t+1}|x_{t+1})$ is then evaluated with the probability distribution of the image patches with respect to the real appearance of the target. Note that the observation process in the appearance-based visual tracking is the warping of an image patch as illustrated in Figure 2.

Because the motion of a target is often unpredictable, and the observation system is usually not explicitly described, the CONDENSATION algorithm (particle filter) [1] is extensively used to implement Bayesian recursions. Even though it is widely used due to its flexibility, however, improvements are necessary to achieve improved robustness and accuracy. In the appearance-based tracking the CONDENSATION algorithm suffers from loss of tracks caused by the approximation error of observation likelihood.

3. INCREMENTAL VISUAL TRACKING

In computer vision applications, dealing with high dimensionality of state vectors and accurate calculations of the observation likelihood are very important but difficult issues in observation system design. In the local coordinate based appearance model, we have a six-dimensional state vector; therefore, it is almost impossible to construct a true observation likelihood distribution.

To deal with the high dimensionality of appearance, we adopt

Figure 3. Filter distraction in IVT

the incremental PCA subspace learning method [5] for template learning. In IVT [5], the observation system is expressed as

$$z_t = h\left(I\left(w\left(x_t\right)\right)\right)+v_t \quad (2)$$

where $w(x_t)$ describes the warping function of the given image $I \in \Re^{N_I \times N_I}$ at the center pixel location of coordinate $p=(p_x, p_y)$ of template (Figure. 1), h is a real-valued function using the cropped image patch as an input argument and, v_t is normal observation noise. In the incremental PCA based observation function, we calculate the mean and M principal eigenvectors and incrementally update them for the reference template appearance. As such, if we let $\bar{T}(x_t)$ and $g_i(x_t)$, $i=1,...,M$, denote the

template mean and M principal eigenvectors, we can represent the reconstruction error matrix for $I\left(w\left(x_t\right)\right)$ as

$$e^2 = \left\| I\left(w\left(x_t\right)\right) - \sum_i c_i g_i\left(x_t\right) \right\|^2, \quad (3)$$

where $c_i = \sum_{x_t} g_i\left(x_t\right)\left(I\left(w\left(x_t\right)\right) - \bar{T}\left(x_t\right)\right)$ are the coefficients from the projection of the template mean to each principal eigenvector $g_i\left(x_t\right)$. IVT has been proved robust when there are time varying changes in object appearance.

However, the PCA based approach cannot reflect fast appearance changes due to the linear structure. Also, the incremental PCA based appearance model tends to experience a filter distraction and a temporal loss of track when there is an abrupt pose change in the object. To effectively adjust to sudden changes in appearance, a small batch size is required for PCA though temporal occlusion needs a larger batch size. Moreover, an adaptive change of the batch size does not offer a complete solution.

4. OBSERVATION LIKELIHOOD APPROXIMATION VIA SIMULATED ANNEALING

4.1 Observation likelihood approximation

The filter distraction is a challenging task in the high dimensional state space because it is very difficult to efficiently identify the high likelihood region. In the appearance-based tracking, if the true appearance is not precisely approximated in the observation system, the filter will gradually adapt to the non-target or the estimate becomes biased. Therefore, many researchers have attempted to design a robust and accurate observation function [6, 4, 7].

To investigate an efficient way of evaluating the observation likelihood in a high dimensional space, the layered sampling approach has been suggested in 3D articulated body motion tracking [7]. This process uses a set of stages to search for the high likelihood region of the state in the observation likelihood distribution. To this end, the annealing process has been introduced so that the most probable state can be locally identified in an efficient manner. In the cascade particle filtering [4], a type of observation evaluation process is proposed that had several different kinds of observers. Sparse and detailed features represent the respective earlier and later stages of weighting functions in layered sampling. In this method, each observation likelihood stage is evaluated using a different classifier, each with a different life span.

Other approaches for solving the filter distraction problem have been suggested in literature. In the MIL [3] approach, authors consider bags of positively and negatively cropped samples (appearances) then tracks the object based on an online boosting algorithm. In [6], visual constraints were considered for penalizing samples of mis-aligned image patches by using an SVM crop classifier so that the filter avoids distraction; our tracker uses an annealing process for observation likelihood approximation.

4.2 Simulated annealing

As discussed in section 4.1, filter distraction problem can be avoided by designing accurate observation likelihood function and finding its maximum. Due to the high dimensional state space

Figure 4. Illustration of the simulated annealing (3 stages)

(given image 240 by 320 pixel in Figure 3), finding out maximum is very difficult job. Fortunately, the region of interest can be found from the learned subspace so that the remaining job is to obtain the best cropped image patch which is best matched with the reference template. In original IVT, there is no such a mechanism so it takes time to recover from the filter distraction or in the worst case the filter drifts. As shown in Figure 3, when an object suddenly changes its pose, IVT temporarily experiences distraction.

To solve this problem we propose to use the well-known optimization method: simulated annealing. In Figure 4, 3 layered weighting functions are displayed to show how to find the maximum in the state space. Distributions in the stage are slightly different from each other. At the beginning the initial distribution is very broad so as to cover overall search space while the last stage distribution is very peak to extract maximum. The relation between weighting function for l stage $\pi_l\left(Z, X\right)$ and the original function $\pi\left(Z, X\right)$ is given as

$$\pi_l\left(Z, X\right) = \pi\left(Z, X\right)^{\beta_l}, \quad (4)$$

for $1 > \beta_0 > \beta_1 > \ldots > \beta_L$, and $l = 1, \ldots, L$ is the stage index. With annealing, the layered sampling can efficiently determine a best cropped sample and prevent the filter distraction.

4.3 Accurate observation likelihood evaluation from annealing process

Basically, in the IVT framework the observation likelihood is evaluated through Sum of Squared Error (SSE) between the reference template and each sample. SSE is represented with the equation (5).

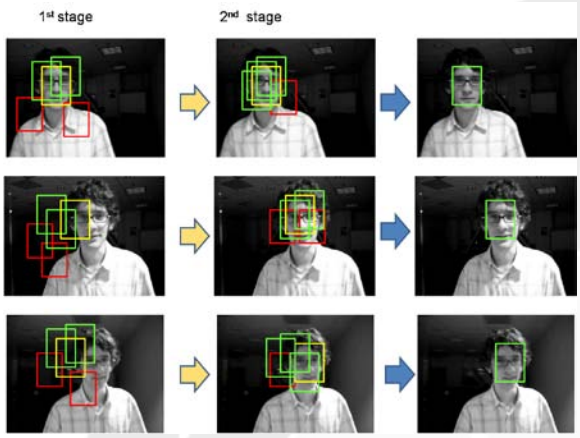


Figure 5. Illustration of the proposed algorithm in visual tracking (yellow box: best matched sample, green box: slightly mis-aligned samples, red box: bad samples

$$SSE\left(x(n)\right)=\sum_{i,j\in M}\left\|I\left(w\left(x(n)\right)\right)-\overline{T}\right\|^2, \quad (5)$$

where, n is sample index, i and j are pixel index of the image patch. Basically, weight for each sample in PF is calculated by using the normalized SSE. With SSE, proposed annealing-based observation function is provided in Overall algorithm.

The whole procedure of the proposed method is summarized in 'Overall algorithm'.

Algorithm 1. Observation process using simulated annealing

for l = 1,2,…L (number of stages)

for n = 1,2,…N (number of samples)

- $x_{k,l}(n) \sim N\left(x_{k,l}; x_{k,l-1}, \pi_l\left(z_k, x_{k,l-1}\right)\right)$

: draw samples from Gaussian distribution for each stage

- Obtain the cropped image patches $x_{k,l}(n)$ from (3)

- $SSE\left(x_{k,l}(n)\right)=\sum_{i,j\in M}\left\|I\left(w\left(x_{k,l}(n)\right)\right)-\overline{T}\right\|^2$

: evaluate SSE for each image patch sample

End for

$\hat{x}_{MAP,l}=\arg\min_x SSE\left(x_{k,l}(n)\right)$

: obtain the MAP estimate for current sampling stage

$x_{k,l+1}=\hat{x}_{MAP,l}$

If $\min\left(SSE\right)<Threshold$

Terminate sampling

Else

Continue

End if

End for

$\hat{z}_k=h\left(I\left(w\left(\hat{x}_{MAP,L}\right)\right)\right)$

: decide the current estimate of the state as the estimate obtained from the last stage

5. EXPERIMENTAL RESULTS

By incorporating termination condition (SSE threshold), 2 or 3 layers are enough to adapt correct template observation based on the experimental results. In the initial stage, a set of samples are drawn from Gaussian distribution of relatively large variance which means sparse sampling. In the consecutive stage, annealing is reflected in the reduced variance of the Gaussian distribution. As illustrated in Figure 5, from a series of annealing process, the best cropped sample with the minimum value of SSE is chosen as the best observation of the current appearance of target.

We test the proposed appearance observation model with real video sequences. Two video sequences are taken from [5] for the evaluation of proposed method. In 'david' sequence, moderate illumination changes and pose variation are incorporated in the appearance of object. From Figure 6, in IVT simulation, filter distraction occurred when the object changes the pose around #161~#181. Even if the IVT experiences distraction, after several sequences it recovers due to the flexibility of PF.

However, when the observation that is best matched to the reference template is not chosen properly in severe situations then the filter gradually adapt toward the non-target. To verify this, more difficult test sequence 'sylvester' is used for performance evaluations of IVT and the proposed method. In 'sylvester' sequence, abrupt pose change is severe so that the filter distraction eventually leads to the tracking failure. Whereas the proposed method effectively handles the abrupt pose changes as shown in Figure 6.

Figure 6. Performance comparison results (green: proposed method, red: IVT)

6. CONCLUSION

A novel observation function is proposed for robust appearance-based tracking. The most challenging issue of filter distraction is tackled by using simulated annealing method in observation function evaluation. From simulated annealing, the best observation (image patch sample) is efficiently selected in high dimensional state space. From the experimental results, the filter distraction due to sudden motion and illumination changes is alleviated by applying proposed method.

7. ACKNOWLEDGEMENT

This work was supported partly by the Basic Research Project through a grant and by the Systems biology infrastructure establishment grant provided by GIST in 2010.

8. REFERENCES

[1] M. Isard, and A. Blake, Condensation-conditional density propagation for visual tracking, *IJCV*, 1998.

[2] J. Kwon, K. M. Lee, and Frank C. Park, Visual tracking via geometric particle filtering on the affine group with optimal importance functions, In *Proc*. CVPR, 2009.

[3] B. Barbenko, M.-H. Yang, and S. Belongie, Visual tracking with online multiple instance learning, In *Proc*. CVPR, 2009.

[4] Y. Li, H. Ai, T. Yamashida, S. Lao, and M. Kawade, Tracking in low frame rate video: A cascade particle filter with discriminative observers of different life spans, *IEEE Trans. on PAMI*, 30(10), 1728-1740, 2008.

[5] D. R. Ross, J. Lim, R.-S. Lin, M.-H. Yang, Incremental learning for robust visual tracking, *IJCV*, 2008.

[6] M. Kim, S. Kumar, V. Pavlovic and H. Rowley, Face tracking and recognition with visual constraints in real-world videos, In *Proc*. CVPR, 2008.

[7] J. Deutscher, A. Blake, and I. Reid, Articulated body motion capture by annealed particle filter, In *Proc*. CVPR, 2000.

Bayesian Network and Minimax Algorithm in Big2 Card Game

Nur Ulfa Maulidevi
School of Electrical Engineering and Informatics ITB
Jl. Ganesha 10 Bandung
+62 22 2508135
ulfa@stei.itb.ac.id

Hengky Budiman
School of Electrical Engineering and Informatics ITB
Jl. Ganesha 10 Bandung
+62 22 2508135
hengky_budiman86@yahoo.com

ABSTRACT
Big2 is a card game that is very popular in East Asia and South East Asia. Strategy and probability to win this card game is very complex. Compare to chess and other board games, research of Artificial Intelligence algorithm to be implemented in card games such as Big2, is infrequent. For that reason, this paper proposed an approach that can be utilized in solving card games problem, especially Big2. There are three candidates that can be explored to solve Big2, Bayesian Network, Minimax, and Rule Based Systems. Based on the comparison of those three candidates and characteristics of Big2 card games, combination of Bayesian Network and Minimax algorithm is chosen to optimize the winning probability. The approach has been implemented in Java platform. Experiment shows that this approach has 23% higher probability compare to average winning probability of human player and compare to the approach using Greedy algorithm.

Keywords
Big2, Artificial Intelligence, Bayesian Network, Minimax, Java.

1. INTRODUCTION
Big2, one of the card games that is very popular in East Asia and South East Asia, is played by four (4) people. The rules are similar to Poker card game, but there are some differences [1]. Each player will have 13 cards, and the one who has empty card first, is the winner. Although there are an online Big2 card game that is played by around 10,000 players, there is not any game application to learn and to practice Big2, as we can find for Bridge [2]. Other advantage of having Big2 application with artificial intteligence (AI) approach is that we can conduct experiments to test other algorithms or approaches for card games. This will give benefit in research of advance searching strategy for intelligent agents.

There are several artificial intelligence approaches that can be implemented to solve Big2, such as Bayesian Network, Minimax algorithm, and Rule Based System. Using Bayesian Network, the causal relationship between variables in the model can be shown [3]. From the relationship, we can calculate the probability of certain variable occurence depends on other variables probability. In Big2, the causal relationship represents card dombination of a player and the action he/ she must conduct.

Minimax algorithm is the basic approach that is used by some of AI based games. This algorithm generates all of the possibilities for every step of each player. The possibilities are represented in decision tree model [4]. In Big2, we need a vast amount of resources to store every possibility of players' step, until the end of the game.

Rule Based System models the way human plays Big2. The inference process inside is used to draw a conclusion, or the choose an action in certain domain [5]. We can exploit this approach to imitate the process of Big2 expert plays the game.

The problem is, we should choose the best approach to be implemented as Big2 solver. The three approaches have advantages and disadvantages. This paper analyzes the best approach, or combination of the approaches, based on the characteristic of card games. The choosen approach is compared to Greedy approach to analyze its performance. Greedy algorithm is often used as comparison to other complex algorithm.

2. BIG2 CARD GAME
This paper only discuss on variant of Big2 card game. This game is very much like poker game. There should be 4 players. Each player receives 13 cards in the beginning. The winner of this game is the first player who does not have any card left. Other players try to minimize their lost by having as minimum cards as possible while other player finishes.

There are three card formats that can be used in Big2, which are:

a. Single

Card format that consists of one card. The cards ranking from high to low: 2 > A > K > Q > J > T > 9 > 8 > 7 > 6 > 5 > 4 > 3. The order of suits from high to low: spades, hearts, clubs, diamonds.

b. Pair

Pair is the combination of two cards, which have the same value of different suits.

c. Packet

Packet is a format that consists of five cards. There are five types of this format, which are: straight, flush, full house, four of a kind, straight flush

The game is started by the player who has 3diamond, and the format he plays must consist of 3diamond. Each player has 13 cards at start. Each game divides into several ticks. A tick is a cycle the players play one format. Each tick is started by the player who wins the previous tick, except for the first tick which is started by the player who has 3diamond. When a player does not have a format to play, he has to pass. A tick is finished when all of the players pass. The winner of this game round is the first player who does not have any card left.

The score of this game depends on the number of cards has been played by the player. For each card, a player receives 1 point. The game is over after several rounds or after certain points has been achieved, based on the arrangement made before the game started. Player who has highest points is the winner.

3. ANALYZES OF THE BEST APPROACH

In general, the Big2 game application requires two capabilities which are:

a. Able to predict the opponent's card;
b. The application able to use the prediction and the card he has to decide the best move, which can increase the probability if winning the game round or minimize the probability of losing.

Based on those two capabilities, analysis of the advantages and drawbacks of Bayesian Network, Minimax algorithm, and Rule Based System is carried out.

3.1 Bayesian Network

Bayesian Network is able to model the probabilities of opponent's cards based on the moves he has played. With careful modeling of relationship between opponent's movements, this ability satisfies the first capability required.

Bayesian Network modeling also easier to understand since it shows the causal relationship between evidence during the game playing. With this structure, the variables that are conditionally independent can be exploit which will be very useful in computation efficiency [3].

Nevertheless, Bayesian Network model does not have the capability to optimize player's movement based on the probabilities it has. We can predict the best movement of a player, but we need to have a very complex Bayesian Network model [6]. The model must represents the probabilities of every player and able to compute the probability of winning in each step. This ability does not suit to the nature of Bayesian Network.

3.2 Minimax Algorithm

Minimax algorithm is usually used to find the optimal move in game application, based on the opponent's movement [6]. In Big2 card game this approach can be exploited to predict the optimal movement of a player.

Although Minimax algorithm is widely used by AI game developer, there are some drawbacks from Minimax algorithm with respect to Big2 card game. The first weakness is that Minimax Algorithm is not able to predict the opponent's cards based on the movement he played. The capability that it has is predicting the combination of cards by assuming that every player has the same probability to have certain card. The computation to produce probability does not include heuristic or observation of a player's strategy.

The second drawback of Minimax algorithm is that it needs a very large search space [6]. For Big2 card game, card combination of 13 cards is more than 40 possibilities. The nature of the game that is partially observed makes it more difficult to model it using Minimax algorithm only. Since each player does not know other players' card, it has to model all the possibilities which affect the search pace. In the worst case, the search tree has 193 in depth and each node has two children nodes in average. In this case, we have to model a search tree which has $1{,}2554*10^{58}$ nodes.

The third drawback of Minimax algorithm is that it assume each player has the same information and belief [4]. This assumption cannot be true since each player can have different information or belief. Minimax algorithm also assumes that each player has the capability of playing the optimal movement, which is also, cannot always be true. Each player has different background and capabilities in playing Big2. Human player is usually gambling and having the best result or the worst result. In Minimax algorithm, each player is assumed to choose the safest move and have in the middle result.

3.3 Rule Based System

Rule Based System can be exploited to model Big2 card game, by representing the knowledge of Big2 expert in rule based. The expert system is used to predict opponent's card by imitating the way an expert predict other players' cards. With this approach, the first capability of Big2 card game can be implemented in Rule Based System.

The drawback of this system is that it will not be able to reason in partially observed environment [5]. There are certain cases where the reasoning conducted in Rule Based System can yield wrong conclusion, when the knowledge is not carefully modeled. For that reason, the design of knowledge representation of Big2 card game in Rule Based System is more difficult and requires more resources compare to Bayesian Network.

Another drawback of Rule Based System is similar to the disadvantage of Bayesian Network. Rule Based System is not able to choose the optimal move based on the prediction of opponents' cards. The large number of possibilities enforce it to have large number of rules to be stored in the knowledge base [5].

3.4 Comparison

Based on the characteristics of each approach discusses in previous section, this section provides summary of the comparison between the three approaches.

From Table 1, the conclusion is that we can not exploit just one approach to implement Big2 card game with capabilities explained in the previous section. In this situation, we can implement two out of the three approaches. The possibilities are combination of Bayesian Network + Minimax algorithm, and Rule Based Sysem + Minimax algorithm.

In this paper, Bayesian Network + Minimax algorithm is choosen since Rule Based System requires an expert of Big2 card game to build the system. The expert is difficult to find, and without the expert we will not be able to verify and validate the knowledge base. Although Minimax algorithm is the only approach that has ability to choose the optimal move, there are some drawback of Minimax algorithm that has to be addressed. In the next section, the solution and the implementation of the choosen approaches are discussed.

Table 1. Comparison of the three approaches

Dimension	Bayesian Network	Minimax Algorithm	Rule Based System
Ability to predict opponents' cards	Yes	No	Yes
Ability to choose the best move	Not efficient	Efficient	Not efficient
Easy to model the Big2 card game	Clear and easy	Clear and easy	Difficult
Processing time	Fast, by exploiting	Slow	Fast, similar to Bayesian

	the causal relationship		Network

4. IMPLEMENTATION

The combination of Bayesian Network and Minimax algorithm needs to be arranged. The arrangements are as follows:

1. Specification of Bayesian Network

This approach is used to predict the opponent's card based on the movement he has made. The prediction includes:

- Possibility that the opponent has the *packet* format;
- Possibility that the opponent has the *pair* format;
- Possibility of the smallest *pair* value the opponent has;
- Possibility of the smallest *single* value the opponent; and
- Possibility that the opponent will have another pass after the last pass he made, for the same combination card format with the smaller value.

2. Specification of Minimax algorithm

The specification of Minimax algorithm is as follows:

- The algorithm is used to choose the best movement based on the card a player has, and the prediction of opponents' card.
- The algorithm uses the Bayesian Network output and calculate based on the probability of each card format combination.

In Big2 card game, usually a player is not able to predict the exact value of the opponents' card. A player can only predict that the opponent has bigger or smaller value compare to his card. For that reason, modeling in Bayesian Network needs certain card classification. Every format combination of cards is categorized into three values, which are *control*, *moderate*, and *straggler*. *Control* is a card combination which usually has a big value and has the possibility to make other players pass in their turn. *Moderate* is a card combination which usually has the medium value. *Straggler* is a card combination which has a small value and can be defeated by other players.

The categorization for each card format combination is as follows.

1. *Single*
   - Control: card 2 for every suit
   - Moderate: card Queen, King, and Ace for every suit
   - Straggler: other cards except the cards in previous categories
2. *Pair*
   - Control: *pair* cards with value Ace and 2
   - Moderate: *pair* cards with value Jack, Queen, and King
   - Straggler: other cards except the cards in previous categories
3. *Packet*
   - Control: combination of *straight flush*, *four of a kind*, and *full house* which is started by Jack as the threes
   - Moderate: combination of *full house* which is started by 3 as the threes.
   - Straggler: combination of *straight* and *flush*

4.1 Bayesian Network Implementation

Based on the specification in previous section, the Bayesian Network is modeled. The structure can be built using either from observation, logic, and experiences; or learning from data. This research uses the first approach, since [6]:

1. Human usually provides a better Bayesian Network structure compare to computer. Computer usually better in calculating the joint probability distribution.
2. Structure obtained from machine learning usually not effective and hard to understand.

There are two types of variables (nodes in Bayesian Network), which are: observed variables and hidden variables. Observed variables are variables that can be perceived by the intelligent agent in this application. This type of variables is represented by single line circle. The hidden variables are variables that cannot be perceived by the intelligent agent. This type of variables is represented by double lines circle. The hidden variables are variables that will be predicted by the application.

There are several Bayesian Networks for Big2 problem modeling. The explanation in this section is only for *pass* condition. The evidences when a player is pass:

1. "Tidak bisa jalan": evidence where the player does not have card combination with correct format to play, and the value is *true* or *false*.
2. "Nilai kartu": evidence which shows the value of existing cards, the value is *control*, *moderate*, or *straggler*.
3. "Jumlah kartu sendiri": the number of cards the player has, the value is *many*, *moderate*, or *few*.
4. "Nilai kartu sudah paling tinggi": an evidence where the cards played has the highest value which is impossible to compete, the value is *true* or *false*.
5. "Simpan kartu": evidence where the player has the card combination to play, but the player choose not to play; the value is *true* or *false*.
6. "Posisi terakhir": evidence where the player who plays at the moment is right before the agent turn; the value is *true* or *false*.
7. "Jumlah kartu lawan": smallest number of opponent's card, the value is *many*, *moderate*, or *few*.
8. "Digunakan untuk format lain": evidence where the player use the existing card for other combination, the value is *true* or *false*.
9. "*Control* format ini": evidence where the player has the *control* value for the format played, the value is *true* or *false*.

Figure 1 depicts the structure for *pass* condition. The explanation of structure in Figure 1 is as follows. The possibility that a player cannot defeat the current card combination is based on the value of the current card combination, the number of cards he has left, and whether the current card combination has the highest value. The possibility that a player keeps a card combination that he can play is based on possibility that he does not have higher value, the position of the last player moved, the minimum number of opponents' card. There are two reasons why a player wants to keep certain card combination; the card could be played as *control* combination, or the card could be played for other combination.

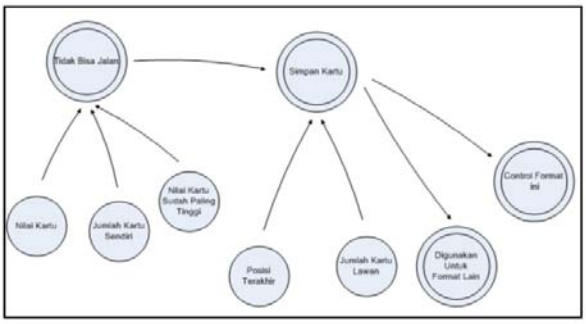


Figure 1. Bayesian network for *pass* condition

An example of inference process from the structure in Figure 1 is as follows. A player passes his turn, when a current card combination is struggler, he still has many cards in his hand, and the value of the current combination card is not the highest value, then the probability of "Tidak bisa jalan" decreases. From that evidence, the position of the last player is just before the agent, and the opponents have many cards, then the probability "Simpan kartu" increases. This evidence also affect the probability of "*Control* format ini" and "Simpan untuk format lain", which will increase as well. The illustration of this inference process is depicted in Figure 2.

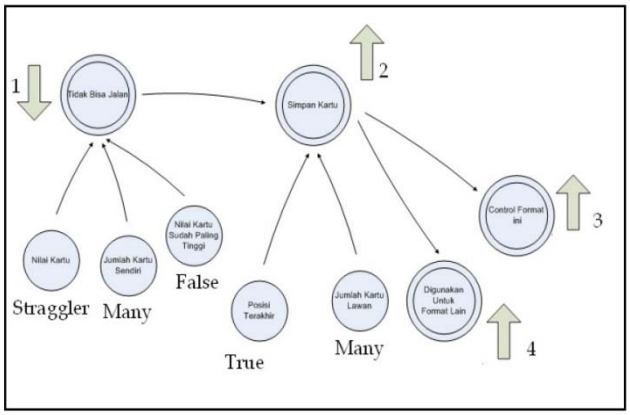


Figure 2. Probability dynamics in *pass* condition

Since Bayesian Network consists of structure and numerical value (joint probability distribution for every node), the conditional probability table (CPT) for the structure in Figure 1 must be completed. A node which does not have parents has the CPT when the value is *true*, while CPT of a child node has the value when it is true for all combination of its parents' value. Only nodes with double line that have CPT. Since the nodes with double lines is hidden from agent's sensor, the CPT of this nodes must be available. The CPT number of these double lines nodes is the input for Minimax algorithm. The value of nodes with singular line can be observed by the agent, and for that reason we do not need to compute the CPT.

The CPT for every node (e_i) is computed using the formula:

$$e_i = \frac{number\ of\ the\ node\ has\ true\ value}{number\ of\ evidence\ for\ the\ node's\ parent} \quad (1)$$

Table 2. CPT of "Tidak bisa jalan" for *single* type

Card value	Number of agent's card	Current card has the highest value	Number of observation	Frequency that the player cannot move	Value
Control	Many	T	20	0	1
Medium	Many	T	20	0	1
Straggler	Many	T	20	0	1
Control	Medium	T	20	0	1
Medium	Medium	T	20	0	1
Straggler	Medium	T	20	0	1
Control	Few	T	20	0	1
Medium	Few	T	20	0	1
Straggler	Few	T	20	0	1
Control	Many	F	20	14	0.70
Medium	Many	F	20	8	0.40
Straggler	Many	F	20	3	0.15
Control	Medium	F	20	18	0.90
Medium	Medium	F	20	10	0.50
Straggler	Medium	F	20	1	0.05
Control	Few	F	20	18	0.90
Medium	Few	F	20	11	0.55
Straggler	Few	F	20	1	0.05

Table 3. CPT of "Tidak bisa jalan" for *pair* type

Card value	Number of agent's card	Current card has the highest value	Number of observation	Freq that player cannot move	Value
Control	Many	T	20	0	1
Medium	Many	T	20	0	1
Straggler	Many	T	20	0	1
Control	Medium	T	20	0	1
Medium	Medium	T	20	0	1
Straggler	Medium	T	20	0	1
Control	Few	T	20	0	1
Medium	Few	T	20	0	1
Straggler	Few	T	20	0	1
Control	Many	F	20	18	0.90
Medium	Many	F	20	12	0.60
Straggler	Many	F	20	5	0.25
Control	Medium	F	20	18	0.90
Medium	Medium	F	20	10	0.50
Straggler	Medium	F	20	5	0.25
Control	Few	F	20	19	0.95
Medium	Few	F	20	13	0.65
Straggler	Few	F	20	7	0.35

Observation of Big2 played by human players must be conducted. The CPT of Bayesian Network for *pass* condition is computed for nodes: "Tidak bisa jalan", "Simpan kartu", "Digunakan untuk format lain", and "*Control* format ini". For each node, we need three CPT for *single* type, *pair* type, and *packet* type. CPT for "Tidak bisa jalan" node for single type is depicted in Table 2, and for pair type is illustrated in Table 3.

The CPT for nodes "Simpan kartu", "Digunakan untuk format lain", and "*Control* format ini" also computed in the same operation, through observation and get the probability.

Other structures with its CPT for double lines nodes that have been modeled for Big2 card game are as follows: Bayesian Network to defeat other player's card (Figure 3), Bayesian Network when an agent starts a new tick (Figure 4), and Dynamic Bayesian Network for pass condition (Figure 5).

For every node that is modeled as double lines circle, the CPT is computed using the same principal and computation as depicted in Table 2 and Table 3.

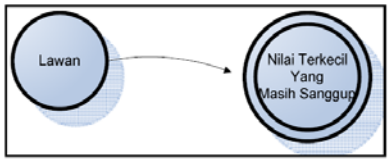


Figure 3. Bayesian network to defeat other player's card

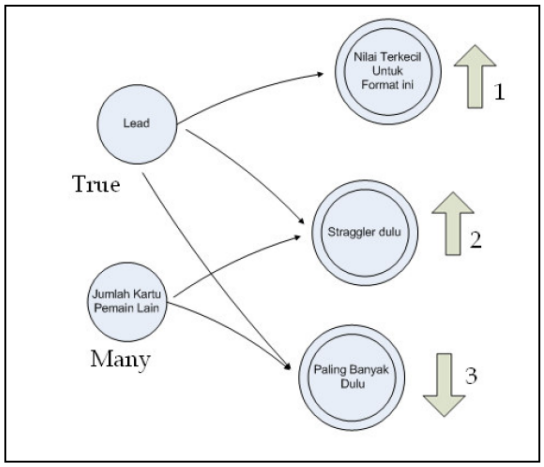


Figure 4. Bayesian network when an agent start a new tick

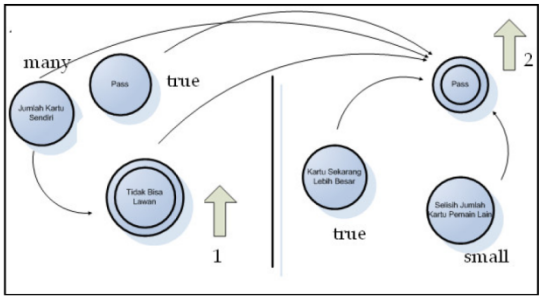


Figure 5. Dynamic bayesian network for *pass* condition

4.2 Minimax Algorithm Implementation

The Minimax algorithm is used the probability resulted from Bayesian Network to choose the best move in order to win the game. The minimax algorithm used in this research is Max^N for multiplayer game. As mentioned in the third section, Minimax algorithm has drawback in large search space. For that reason, the simplification of the search three is urgently required.

Alpha-beta pruning is not feasible in this approach, since there is no value balancing in this game. The simplifications that is implemented in this approach are:

1. The game three is limited until 9 depths. This level is enough to predict three steps ahead, and still has the reasonable search space.
2. The limitation is not useful if it is built in the beginning of the game, since the Bayesian Network can not provide enough information, and there are numbers of possibilities to be considered. For that reason, Minimax algorithm is used in the middle of the game. In the beginning of the game, the players exploit heuristic to choose the best move without Minimax algorithm. The heuristic implemented in this research is Greedy algorithm.
3. When computation in Bayesian Network provides high probability for certain evidence, the Minimax algorithm expands only the branch that has the evidence. Other branch in search three will not affect much in choosing the optimal move, since the probability is very small.

The problem arises when there is a missing value in Bayesian Network, and the value is needed by Minimax algorithm. The solution of this problem is by using 'default' value for the missing value in Bayesian Network. The default value is obtained from observation, that is carried out during the making of Bayesian Network structure and CPT value.

The example of Minimax usage in Big2 card game is as follows.

1. Player A has 10♦, 10♣, 10♠, A♥, A♠.
2. Player B has a card left.
3. Player C has player C has 10 cards left, and information from Bayesian Network is that the probability that player C will *pass* again for bigger *pair* value than pair 5 is 0.925
4. Player D has 3 cards left and now is leading.

Player D starts the tick with pair 7. Agent will produce search three with Minimax algorithm as depicted in Figure 6.

The square represents agent's (player A) movement, the diamond represents player B's movement, the triangle represents player C's movement, the circle represents player D's movement. From the tree in Figure 6 player A will choose to play pair 10 because it is considered as the best move.

The implementation of the application uses Java language of IDE NetBeans 6.0 in Microsoft Windows XP environment. There are four players shown in the application, player 1 is an agent using the proposed approach, player 2 using Greedy, player 3 using Greedy, and player 4 is a human player. The application board is depicted in Figure 6.

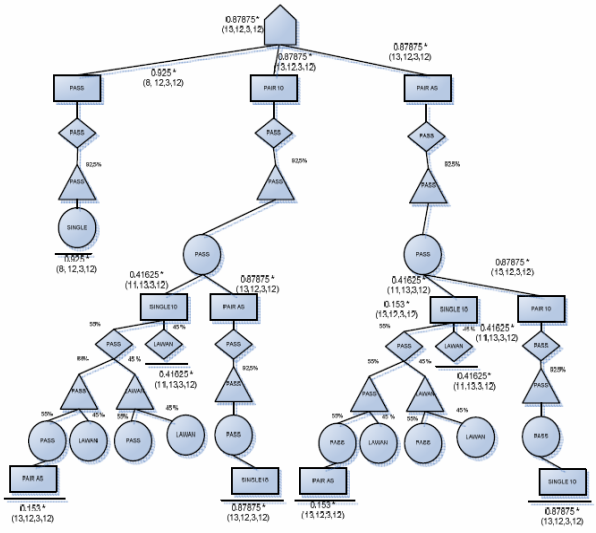


Figure 7. Game tree in Big2 card game

Figure 8. Application board interface

5. EVALUATION

Evaluation is conducted to figure out the performance of the selected approach, which is Bayesian Network + Minimax algorithm (BNM) for Big2 card game. The evaluation consists of 4 games, each game has a player that implements BNM, two players that implements Greedy, and a human player who has 2 until 6 years experience in playing Big2.

The frequency of a player winning, in second place, third place, or fourth place is illustrated in Table 4. The point in each round collected by each player is depicted in Table 5.

In each round, the winner received 4 points, the second place received 3 points, the third place received 2 points, and the last place received 1 point. In average, probability of each player's point is 2,5. Since this evaluation is conducted in 40 rounds, the average point of each player is 100. BNM shows that it has 123 point, 23% higher then the average, and it has highest winning frequency and points compare to greedy algorithm and human player.

Table 4. Frequency of winning

Ranking	BMM	Greedy1	Greedy2	Human
1	3	0	0	1
2	1	0	1	2
3	0	1	2	1
4	0	3	1	0

Table 5. Statistics in each round

Position	BNM	Greedy1	Greedy2	Human
1	18	6	7	9
2	11	11	8	10
3	7	12	13	8
4	4	11	12	13
Total points	123	92	90	95

6. CONCLUDING REMARKS

From the evaluation, Bayesian Network and Minimax algorithm have better performance compare to greedy algorithm and human player. This result increases the confidence that this approach is feasible to be implemented in card game, and as a benchmark approach in searching strategy.

This is a preliminary research in applying Bayesian Network together with Minimax algorithm for card game. There are many things that could be carried out to improve the solution quality and performance, such as adding more variables in Bayesian Network model, changing card category based on the format of current card, and optimization of Minimax algorithm.

7. REFERENCES

[1] Tai, Kang Meng, et. al. http://www.pagat.com/climbing/sps2172.pdf

[2] -, Viwawa. http://www.viwawa.com/en_US/game/Big25

[3] -, An Introduction to Bayesian Network and Their Contemporary Applications. http://www.niedermayer.ca/papers/bayesian/bayes.html

[4] Sturtevant, Nathan Reed. (2003). Multi-Player Games: Algorithms and Approaches. University of California

[5] Paine, Jocelyn. What Is A Rule Based System. http://www.jpaine.org/students/lectures/lect3/node5.html

[6] Russel, Stuart, et. al. 2003. Artificial Intelligence: A Modern Approach. Prentice Hall.

Cell Formation Using Particle Swarm Optimization (PSO) Considering Machine Capacity, Processing Time, and Demand Rate Constraints

Dedy Suryadi
Universitas Katolik Parahyangan
Ciumbuleuit 94, Bandung 40141
Indonesia
62-222032700
dedy@home.unpar.ac.id

Ferry Putra
Universitas Katolik Parahyangan
Ciumbuleuit 94, Bandung 40141
Indonesia
62-222032700
pice_boy2003@yahoo.com

Cynthia Juwono
Universitas Katolik Parahyangan
Ciumbuleuit 94, Bandung 40141
Indonesia
62-222032700
juwonocp@home.unpar.ac.id

ABSTRACT
Group Technology (GT) layout is a layout that combines advantages from product layout and process layout. Cell formation is the first step to apply it. In this research, Particle Swarm Optimization (PSO) algorithm is used for machine grouping, while degree of belongingness is used for grouping part families. The considered constraints are machine capacity, processing time, and demand of each part. Performance measure used as the objective function is generalized grouping efficacy.

The model is implemented in 8 cases which have differences in number of machines, number of components, demand of each component, and processing times. Each case is tested with several combinations number of parameter wmax (inertia weight maximum), c1 (self confidence), c2 (social confidence).

The result shows that matrix size affect sensitivity model to parameter. Smaller matrix is not sensitive to PSO parameters. Larger matrix is sensitive to parameter c, where using large c results in difficulty to reach convergence but reach better global best faster than using small c. Parameters c1 = 1 and c2 = 2 show less iterations needed to reach global best. Based on the implementation, it can be shown that the model has better performance compared with Tabu Search and Genetic Algorithm.

Keywords
group technology, cell formation, PSO, meta-heuristic, layout

1. INTRODUCTION
Cell manufacturing is an application of group technology (GT) concept. It forms a cell which contains a group of machines to produce a group of components having relatively high similarity. Cell formation is done based on the information about the operations needed for each component. It may be represented in a binary matrix, where 1 means a component needs to be processed in a particular machine, while 0 is otherwise.

The typical binary component-machine matrix assumes infinite capacity and abandones processing time, demand rate, and process sequence. It is actually possible that the load for a type of machine is too large to be done by a single machine of that type. The load is defined to be the capacity required by all components to be done on that type of machine. For a type of machine, it should be computed first how many machines of that type is needed. Suppose it is needed more than one machines for that type, those machines may be or may be not allocated to the same cells.

The conventional cell formation algorithms such as *Rank Order Clustering* (ROC), *Direct Clustering Algorithm* (DCA), *Cluster Identification Algorithm* (CIA), *Bond Energy Algorithm* (BEA), do not consider machine capacity, processing time, and demand rate of each component. Thus, another cell formation algorithm is needed to accommodate those constraints.

The conventional performance masures for cell formation such as *grouping efficiency* (η) and *grouping efficacy* (τ) do not consider processing time. The measure that shows a more comprehensive performance of a cell formation is the *Generalized Grouping Efficacy* (Γ_g) which considers processing time and demand rate of each component (Rogers and Shafer, 1995).

This research develops a cell formation model considering machine capacity, processing time, and demand rate of each component, based on *Particle Swarm Optimization* (PSO). The previous researches of the same problem has been done by Tabu Search (Kelly, 2006) and Genetic Algorithm (Zolfaghari and Liang, 1998).

2. RESEARCH OBJECTIVES
The objectives of this research are: (1) Applying *Particle Swarm Optimization* (PSO) algorithm on cell formation considering the constraints of machine capacity, processing time, and each component's demand rate; (2) Investigating the PSO parameters' influence to the developed model's performance; (3) Comparing the developed model's performance with previous models based on Tabu Search and Genetic Algorithm.

3. ASSUMPTIONS and LIMITATIONS
The assumptions taken for this research are: (1) The processing time of each component has included loading, setup, run, and unloading time; (2) Duplicated machines are identical; (3) Demand rate's change over time is insignificant; (4) The capacity required is average daily required capacity; (5) Increasing capacity is done only by adding machine(s).

The problem limitations are: (1) Process routing is not considered; (2) Physical machines layout in a cell is not considered; (3)

Transportation time is neglected; (4) The capacity needed to process each component does not exceed the machine's capacity.

4. GROUP TECHNOLOGY

Group Technology (GT) is a concept that identifies and groups similar components into a family, such that the fabrication process of those families may become more efficient (Tompkins et al, 1996). Cellular manufacturing is an application of GT concept. The basic idea is grouping machines into cell to process particular part families. The objective is such that the components processed in a cell spend most of the time inside its own cell and have minimal interactions with other cells, such that the production efficiency is increased.

The key of a successful cellular manufacturing starts with the good cell formation. The absolute independence between cells is expected, but is relatively difficult to achieve in reality. When a component needs to be processed in cells other than where it actually belongs, it is called exceptional part and the machine processing it is called bottleneck machine. Such intercell movements results in the increase of work-in-process, material handling cost, and the decrease of production efficiency.

Cell formation itself is a clustering problem, which is also a combinatorial problem. Meta-heuristics methods have been more and more used recently to solve such problems.

5. PARTICLE SWARM OPTIMIZATION

PSO is a meta-heuristic algorithm which is inspired by social psychology science. It was firstly introduced by Kennedy and Eberheart (Kennedy et al, 2001). The algorithm is based on social behavior of a flock of bird or a school of fish searching for food, in which they spread and finally converge to a point. The basic principles of PSO are (Kennedy et al, 2001):

1. each particle has a particular position and velocity.
2. each particle knows its position and its respective objective function value.
3. each particle remembers its best position (personal best) and its respective objective function value.
4. each particle knows the best position has been reached so far (global best) and its respective objective function value.
5. each particle is able to follow its environment.
6. in each iteration, the position change of each particle follows three patterns: exploring a new area, moving to the direction of its personal best, moving to the direction of global best.

The steps of PSO algorithm are:

A. Generating initial position and velocity

The initial position and velocity for each particle are generated by the equations:

$$x_0^i = x_{\min} + rand\left(x_{\max} - x_{\min}\right) \quad \text{(Eq. 1)}$$

$$v_0^i = \frac{x_{\min} + rand\left(x_{\max} - x_{\min}\right)}{\Delta t} \quad \text{(Eq. 2)}$$

where:

x_0^i = position of particle i at iteration 0

X_{min} = minimum allowable position of any particle

X_{max} = maximum allowable position of any particle

v_0^i = velocity of particle i at iteration 0

Δt = period of time

B. Updating velocity and position

Velocity and position of each particle are updated continuously at each iteration by the equations:

$$v_{k+1}^i = w_k v_k^i + c_1 r_1 \frac{\left(p_k^i - x_k^i\right)}{\Delta t} + c_2 r_2 \frac{\left(g_k^i - x_k^i\right)}{\Delta t} \quad \text{(Eq. 3)}$$

$$x_{k+1}^i = x_k^i + v_{k+1}^i \Delta t \quad \text{(Eq. 4)}$$

$$w_k = \beta . w_{k-1} \quad \text{(Eq. 5)}$$

where:

v_{k+1}^i = velocity of particle i at iteration (k+1), or to reach position at iteration (i+1)

x_k^i = position of particle i at iteration k

w_k = *Inertia weight* at iteration k

p_k^i = *Personal best* of particle i at iteration k

g_k^i = *Global best* of particle i at iteration k

β = decreasing factor

c_1 = personal confidence rate (believes in itself)

c_2 = social confidence rate (believes in other particles)

r_1, r_2 = random number between 0 and 1

If a particle's position falls outside the feasible region, then its position must be put back into the feasible region. There is a pulling back method introduced (Toyoda et al, 2006), in which:

1. the particle is pulled back to the nearest point inside the feasible region

2. the particle is pulled back to the nearest extreme point of the feasible region
3. the particle is pulled back to the feasible solution and placed randomly inside it

C. Decoding each particle and evaluating the respective objective function

D. Stopping the search based on the stopping criterion, such as number of iterations or particle position's convergence.

6. DEVELOPING CELL FORMATION MODEL

This section discusses the particle, its boundaries, and its performance measure.

6.1 Particle Encoding

The particle in PSO should be encoded uniquely and fully represents a solution uniquely as well. In this model, the particle is encoded using random key representation (Bean in Gen and Cheng, 1997). The next example illustrates the particle encoding used in this model.

Suppose there are 6 machines to be grouped into cells, so the particle contains 6 dimensions. Let there be a particle as follows:

[1.34 1.35 2.56 2.47 3.45 3.12]

The first digit of each dimension is taken as the keys:

[1 1 2 2 3 3]

The keys can further be translated this way. The machines 1 and 2 are placed in cell 1, machines 3 and 4 in cell 2, machines 5 and 6 in cell 3.

Such encoding, however, may result in same solution although the particles are different. The next table illustrates such problem.

Table 1. Keys

No	Keys	Cell 1	Cell 2	Cell 3
1	[12321132]	M1,M5,M6	M2,M4,M8	M3,M7
2	[23132213]	M3,M7	M1,M5,M6	M2,M4,M8
3	[31213321]	M2,M4,M8	M3,M7	M1,M5,M6

To avoid the duplication, the keys are further evaluated. The key of machine 1 (regardless of its value) will be assigned to cell 1. If machine 2 has the same key with machine 1, it will go to the cell 1 (which has been previously formed), otherwise it would form a new cell. The modified keys of the particles are shown in the next table.

Table 2. Modified keys

No	Modified Keys	Cell 1	Cell 2	Cell 3
1	[12321132]	M1,M5,M6	M2,M4,M8	M3,M7
2	[12321132]	M1,M5,M6	M2,M4,M8	M3,M7
3	[12321132]	M1,M5,M6	M2,M4,M8	M3,M7

6.2 Solution Space Boundaries

In this research, each particle jumping out from the feasible region is pulled back into the feasible region by the mirroring principle. The next example should make the principle clear.

Let there be 5 machines to be grouped into cells. There is a particle at iteration k ($k \neq 0$) as follows:

[1.6 2.8 11.5 6.8 3.4]

Those 5 machines at most may be grouped into 5 cells, so the value of each dimension should be less than 5. Therefore it can be seen that 11.5 and 6.8 are outside the feasible region. The 6.8 value is too far 1.8 away from 5, thus it is mirrored back to become (5 – 1.8) = 3.2. Thus, all values between 5 and 10 is mirrored. However, 11.5 is pulled back to become (11.5 – 10) = 1.5, because it has started a new pair of mirrors, i.e. from 10 to 20.

6.3 Required Machines Calculation and Component Assignment

For each type of machines, the capacity required on that type for all components is calculated. If the result is larger than the capacity of a single machine of that type, it means that type of machine must be duplicated until the required capacity is fulfilled. Later, each duplicated machine is treated as independent machines.

For duplicated machines, the component must be assigned to each machine. The assignment is done based on workload. Workload is obtained for each component, in which it is the multiplication of processing time and demand rate of each component. The assingment starts from the largest workload to one machine, then the second largest to the next machine, and so on. The purpose of such assignment is to achieve belongingness of a component to its cell as high as possible. Even if at the end there is an exceptional part, it is expected to be the component having small workload.

6.4 Evaluating Exceptional Part on Duplicated Machines

Once the cells have been formed, exceptional parts may be detected. This step aims to eliminate exceptional parts (if exist) at duplicated machines through a re-assignment of workload as follows. For each exceptional part exists:

1. Find the targeted cell of the exceptional part. The targeted cell is the cell in which the part has the highest degree of belongingness.
2. In the targeted cell, find the type of machine needed by the part. All machines of that type are emptied from previous workload assignments.
3. Sort all workloads from the largest, including the workload of the exceptional part.
4. Assign the largest workload to the first machine. Next, assign the largest possible workload considering the remaining capacity of the first machine. If there are no more workloads able to be assigned to the first machine, go on to the second machine and repeat the assignment step from the largest, and so on.

5. If at last there are workloads unable to be assigned, the corresponding parts must be exceptional parts.

6.5 Performance Measure

Generalized grouping efficacy is proposed to measure performance of a machine grouping (Zolfaghari and Liang, 2003):

$$\Gamma_g = \frac{t_d}{t_0 + \sum_{r=1}^{R}(M_r \sum_{j=1}^{N}{}_{j\in\Omega_r} L_j t_j^{\max})} \quad \text{(Eq 6)}$$

Γ_g = *generalized grouping efficacy*

t_d = total processing time inside cell, where $t_d = \sum t_{ij} \times L_j$ and $i \in \Lambda_r \; j \in \Omega_r$

t_0 = total processing time outside cell, where $t_0 = \sum t_{ij} \times L_{j\,r}$ and $i \notin \Lambda_r \quad j \notin \Omega$

M_r = the number of machines in cell r

L_j = demand rate of component j

t_j^{max} = maximum process time of component j

Ω_r = components inside cell r

Λ_r = machines inside cell r

6.6 Degree of Belongingness

To check how high a component's belongingness to a particular cell, the degree of belongingness may be used. The component is assigned to the cell where it has the highest degree of belongingness. The degree Drj = 1 indicates component j is absolutely suitable to be inside cell r, while Drj = 0 indicates component j is absolutely unsuitable to be inside cell r. The formula:

$$D_{rj} = \frac{m_{rj}}{M_r} . \frac{m_{rj}}{m_j} . \frac{h_{rj}}{H_j} \quad \text{(Eq. 7)}$$

D_{rj} = *degree of belongingness* of component j to cell r

m_{rj} = the number of types of machines in cell r required by component j

h_{rj} = processing time of component j in cell r

M_r = the number of machines in cell r

m_j = total number of types of machines required by component j

H_j = total processing time required by component j

. The right and left margins should be 1.9 cm (.

7. CELL FORMATION ALGORITHM

The complete algorithm follows the following steps:

Step 1 :

Determining the inputs:

a. component-machine matrix consists of t_{jm}, i.e. processing time of component j at machine m
b. demand rate of component j, Lj
c. each machine's available capacity per period

Step 2 :

Determining PSO parameters:

a. number of particles, n
b. maximum and minimum inertia weights, w_{maks} and w_{min}
c. particle's confidence to itself, c_1
d. particle's confidence to other particles, c_2
e. maximum interation, K_{maks}

Step 3 :

Comparing machine's available capacity and the workloads. For any type of machine, if the available capacity is not adequate, it must be duplicated until the workloads can be handled. If duplication is performed next go to Step 4, otherwise to Step 5.

Step 4 :

For duplicated machines, components are assigned based on the highest workload. The highest is assigned to the first machine, the second highest to the next, and so on alternatively. The component-machine matrix must be revised, because now some machines have been duplicated and the components have been specifically assigned to each duplicated machine.

Step 5 :

Set K = 0 and i = 1

Step 6 :

Determine the upper bound and lower bound for particle's position values. Generate the initial position and velocity of particle using Eq.1 and Eq.2, respectively.

Step 7 :

Decode particle's position into modified keys, such that the number of cells and machine grouping are obtained. For the initial solutions, each cell must have at least a machine inside it.

Step 8 :

Allocate the components into the cells formed. Each component's degree belongingness is computed with respect to each cell using Eq.7. A particular component is assigned to a particular cell in which it has the highest Drj.

Step 9 :

Evaluate exceptional part on duplicated machines.

Step 10 :

Compute the performance (generalized grouping efficacy) of current particle using Eq.6. At K=0, each particle's performance automatically becomes its current personal best as well.

Step 11 :

Check whether i has reached number of maximum particle (n). If not, set i = i+1 and go back to Step 6. otherwise, go to Step 12.

Step 12 :

Find global best of all particles, i.e. the best generalized grouping efficacy has been reached by any particle so far. At K=0, the best personal best automatically becomes the global best as well.

Step 13 :

Change K into K + 1.

Step 14 :

Change i into 1.

Step 15 :

Update particle's velocity using Eq.3 to move a particle from its previous position (use modified keys) to the current at this iteration. The new position is obtained by Eq.4.

Step 16 :

Decode the particle's position into modified keys and obtain the machine grouping based on them.

Step 17 :

Allocate the components to the cells based on degree of belongingness.

Step 18 :

Evaluate the exceptional part on duplicated machines.

Step 19 :

Compute objective function value of the cell formed. Find the personal best of particle i by comparing the objective function value at iteration K and at iteration K-1. If larger, then the personal best is updated by the current objective function value, otherwise not.

Step 20 :

Check whether i has reached n. If not, update i into i+1, and go back to step 15. Otherwise, go to step 21.

Step 21 :

Among all personal bests at iteration K, find the largest value. If the current global best is less than that value, update the global best with that value. Otherwise global best remains.

Step 22 :

Evaluate all particles' position. If all particles are at the same position, it is said to be at convergence and the algorithm stops. Otherwise, go to step 23.

Step 23 :

Check whether K is less than K_{maks}. If yes, go back to step 13. If not, the algorithm stops.

8. CASE IMPLEMENTATION

The developed algorithm is implemented to cases from previous researches (Kelly, 2006 and Zolfaghari and Liang, 2003) in order to compare the performances with other algorithms.

8.1 Parameters

The parameters used for implementation:

- K_{maks}(maximum iteration) : 5000
- w_{min} (minimum *inertia weight*) : 0.1
- w_{maks} (maximum *inertia weight*) : 0.5, 0.9, 1.4
- c_1 (particle confidence on itself) : 0.5, 1, 2
- c_2 (particle confidence on other particles) : 1, 2, 5
- n (number of particles) : 20

Other determined values :

- $\beta = 0.875$; based on [14]
- $[(X_{min} - X_{max}) / 2]$ for V_{min}
- $[(X_{max} - X_{min}) / 2]$ for V_{max}
- Convergence = 10 % of all particles have different positions.

8.2 Case Description and Comparison

There are 3 groups of case. Group 1 contains cases having relatively small matrix, the data is hypothetical. Group 2 contains cases from Kelly [5]. Group 3 contains cases from (Zolfaghari and Liang, 2003), which is also used by (Kelly, 2006). The grouping result of the model developed in this research is compared with cases in Group 2 and 3.

Table 3. Case comparison

No	Component	Machine	Demand
1a	10	8	Uniform
1b	10	8	Various
1c	10	8	Various
2a	42	10	Uniform
2b	42	10	Various
3a	43	22	Various
3b	40	24	Various
3c	40	24	Various

8.3 Implementation Result and Comparison

It can be generalized that the best combination to obtain better global best is $c_1 = 2$ and $c_2 = 5$. The best combination for average iterations to obtain global best is $c_1 = 1$ dan $c_2 = 2$. The best value for w to obtain global best is 1.4. The performance, in terms of Γ_g, of different models are also compared for cases in Group 2 and 3 as follows.

Table 4. Comparison of Γ_g for Group 2

No	PSO (average)	PSO (std deviation)	TS
2a	0.641	0.008	0.491
2b	0.677	0.021	0.556

Table 5. Comparison of Γ_g for Group 2

No	PSO (average)	PSO (std deviation)	TS	GA
3a	0.549	0.011	0.408	0.397
3b	0.633	0.005	0.619	0.532
3c	0.559	0.020	0.562	0.464

All material on each page should fit within a rectangle of 18 x 23.5 cm (7" x 9.25"), centered on the page, beginning 2.54 cm (1") from the top of the page and ending with 2.54 cm (1") from the bottom. The right and left margins should be 1.9 cm (.

9. CONCLUSIONS

The conclusions of this research are:

1. The PSO-based algorithm for cell formation developed in this research is as shown on Section 7.

2. The influence of parameters to model performance varies by cases:

a. in Group 1, c and w do not influence global best, but c influences the number of iterations to reach global best.
b. in Group 2, w generally does not influence global best and iterations to reach it, while c does.
c. in Group 3, w does not influence global best, while c generally influences global best and iterations to reach it.
d. larger c_1 and c_2 reach better global best but are difficult to reach convergence.
e. there is no influence of interaction between w and c.

3. The comparison with previous researches:

a. in Group 2, the developed model's performance is better than TS-based model of (Kelly, 2006), because Kelly did not evaluate exceptional parts in duplicated machines and the number of cells are fixed.
b. in Group 3, the developed model's performance is better than GA-based model of (Zolfaghari and Liang, 2003) and generally better than Kelly's as well.

10. REFERENCES

[1] Gen, Mitsuo dan Cheng, Runwei., *Genetic Algorithms & Engineering Design*, New York : John Wiley & Sons, Inc., 1997.

[2] Kelly, Wray., *Pengembangan Algoritma Pembentukan Sel Yang Memperhatikan Kebutuhan Kapasitas Dan Syarat Kedekatan Tiap Mesin Menggunakan Tabu Search*, Bandung : Jurusan Teknik Industri UNPAR, 2006.

[3] Kennedy, James., Eberheart, Russel C., Shi, Yuhui., *Swarm Intelligence*, San Fransisco : Morgan Kauffman Publisher, 2001.

[4] Rogers, D.F. & Shafer, S.M., *"Measuring Cellular Manufacturing Performance"*, Elsevier Science, vol 24, pp. 147-163, 1995.

[5] Tasgetiren, M. Fatih, Liang, Yun Chia, Sevkli, Mehmet, dan Gencyilmaz, Gunes, *"A Particle Swarm Optimization Algorithm for Makespan and Total Flowtime Minimization in The Permutation Flowshop Sequencing Problem"*, European Journal of Operational Research, 2006.

[6] Tompkins, White, Bozer, Frazelle, Tanchoco, dan Trevino., *Facilities Planning 2nd ed.*, New York : John Wiley & Sons, 1996.

[7] Toyoda, Y., Shohdoji, T., dan Yano, F., *"An Application of Particle Swarm Optimization to Linear Programming Problems "*, The 36th CIE Conference & Industrial Engineering, 2006.

[8] Zolfaghari, S. dan Liang, M., *"A New Genetic Algorithm for The Machine / Part Grouping Problem Involving Processing Time and Lot Sizes"*, Elsevier Science, Computer & Industrial Engineering, vol 45, pp.713-731, 2003.

Computer Aided Learning for List Implementation in Data Structure

Ng Melissa Angga
Informatics Engineering, University of Surabaya
Kalirungkut Street,
Surabaya, Indonesia
(+62) 31 2981393
melissa@ubaya.ac.id

Susana Limanto
Informatics Engineering, University of Surabaya
Kalirungkut Street,
Surabaya, Indonesia
(+62) 31 2981395
susana@ubaya.ac.id

ABSTRACT
Data Structure is one of the core subjects in most Information Technology Faculty which considered as a hard subject that many students failed to understand the content of this subject. Through some analysis, found out that this problem caused by the lack of student motivation towards this subject, the lack of ability in picturing the process behind this subject, and the failure to comprehend the topic about list which heavily related to many other topics. List topic apparently is a crucial elementary topic in data structure, in which the failure of understanding in this topic would make it impossible to understand the other subsequent topics. In this paper, we offer an alternative way to presenting the topic about list in data structure using multimedia technology.

Keywords
List, Data Structure, Computer Aided Learning

1. INTRODUCTION
Data Structure is one of the subjects which most Information Technology Faculty adapted in their core curriculum. In data structure, students learn on how to organize data, save and manipulate them. Data structure would be the basic knowledge and prerequisite for many other advance subjects in Information Technology.

As much as the important role of Data Structure as a core competency needed for any Information Technology scholar, this subject is apparently not easy to be delivered. The number of failed students of this subject in one university is nearly one third of all members of the class. These issues should be taken seriously since the lack of competencies in Data Structure subject would resulting in the lack of competencies in any other subjects related.

Among many topics covered in Data Structure, topic about List stood up as the fundamental topic which applied and elaborated in many other topics. Having said that, some investigation also shows that student who failed on the topic of List would also failed on other topics.

Based on those facts, this research is conducted to offer a better way to deliver the topic about List in Data Structure subject. From some study, the authors found that multimedia technology is a very powerful tool that can be used to create a clearer and more concise presentation. Thus, multimedia technology would be adapted in this research. However the scope of this research would only limited to the implementation phase only, since the result of this research was not yet applied to the students of Data Structure subject.

2. LIST
Nyhoff and Leestma (1992) identify list as a limited series of elements in data structure. List actually has been used in our every day life, for example people use shopping list to write down items needed to be bought, in the dentist waiting room, the nurse hold a list of the names in which she should call in sequence, a secretary has a list of activities should be done by her boss in sequence today, etc. Each list has some fundamental operations relevant with it, they are listed as follows:

1. Create an empty list
2. Check if a list is empty
3. Traverse the elements of some parts of the list
4. Insert new element to the list
5. Delete elements from a list
6. Check whether a list is full

Since list is described as a sequence of elements, therefore there is an order of elements being placed on the list. There would be the first element, second element, and so on up to the last one. This ordering should be reflected on the implementation of the list.

The easiest way to create a list with an implicit ordering rule is by using an array which by default has its own ordering method.

Order	1	2	3	4	5	6	7
Data	a	a	b	A	z		

Figure 1. Implicit order of a list.

However easiness in creation is not a guarantee for the easiness in maintenance. To insert and delete an element to a list with implicit ordering rule required shifting of many other elements associated. For instance, if one need to delete the second element of the above list example, then the third element should shifted to the second place, the fourth to the third place, and the fifth to the fourth place, then the result would be something like this.

Order	1	2	3	4	5	6	7
Data	a	b	a	Z			

Figure 2. List after deleting the second element

With the same manner, inserting a new element to the second place required shifting of the fourth element to the fifth place, the third to

the fourth place, and the second to the third place, before the new element injected to the second place. Maintaining a list in this way would be inefficient both in time and resource allocation.

To address the problem arise with implicitly ordering list, there is another alternative way to explicitly show the order of a list. This kind of method known as linked list. Each part of a linked list consists of the data and the address of the next element. Thus, if one has located the address of the first element, he would be able to determine every other element consecutively through the end. A representation of a linked list is shown on the next figure.

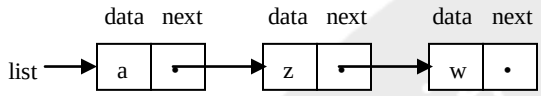


Figure 3. Linked list representation

Linked list data structure can be implemented using array or pointer. In both case, a class should be prepared to save data and the next address.

3. COMPUTER AIDED LEARNING

As been predicted by Baldwin and Down in Education Technology for Engineering (1981), the cost of technology in education process now has been too little to be ever put into consideration. Thus, the issue right now is not whether an educational technology is affordable, but it is how to exploit it in an appropriate manner.

The term Computer Aided Learning, or CAL in short, by itself covers both the educational parts in which the teacher set up and organize some teaching materials, and the technological parts in which a software and a computer used to aid the whole learning process. CAL typically aimed for some ambitions such as to cut down costs (by efficiently decrease the investments for other teaching materials and teacher working hours), to enhance the learning experience by closing gaps between theory and practice, and to serve the broaden coverage area.

Reddi in Educational Multimedia, A Handbook for Teacher-Developers (2003), describe multimedia in our world today as a compilation of text, graphic art, sound, animation and video elements. Whereas an interactive multimedia is a multimedia project which allow the alteration of presentation by the end user, in other words, the end user are able to change 'what' to be presented, 'when' is the time to presented, and 'how' is the presentation.

The popularity of games development nowadays can be used as an indication on how multimedia presentation would be accepted in education world. Multimedia technology offer richer and clearer presentation, moreover the presentation looks better and could gain more interest from the students.

Having said that, multimedia technology offer a better way to simulate process which hard to be presented in traditional way via speech and text only. This simulation would give a better explanation and let it stay longer in the memory.

4. ANALYSIS OF THE PROBLEM

Towards the low percentage of the number of students who fully accomplished the data structure subjects, a study has been made to find out the root problems behind those results. The examination has been conducted through a survey to some previous students. Some conclusion regarding the root cause of the problems has been drawn after the completion of the survey, those conclusions were:

1. Students don't have enough motivation towards the subject. This problem arises because this subject has been gained its reputation as one of the killer subject. Previous alumni used to address this subject as hard and difficult. This kind of addressing would inevitably lower the motivation of the student even before they ever have a touch of the subject itself. And this assumption is aggravated once they step in to the class and learn some early topics of this subject (one of them was the topic about List).
2. Student's background knowledge makes him/her incapable to picture the process behind some topics. Thus they hardly understand the explanation about the process.
3. Some topics are considered prerequisite for other topics. Therefore the failure to handle one topic can devastate the chance to understand another topic. One topic related to this matter is List.

5. DESIGN AND IMPLEMENTATION

Based on investigation on the recent learning activities as stated above, a design for new presentation of "List" has been proposed. In addressing the lack of motivation problems, this presentation should be interesting enough and promote the clear concise explanation regarding the topic. Considering the difficulty of the student to picture the background process of the topic, this presentation should provide a step by step simulation of the process.

The presentation of the List topic is started by the general explanation of list. After a brief introduction with the list, then the student would be conducted through two different paths one at a time. The first part is the presentation of the sequential storage for the implicit presentation of list order. The next part is created as an explanation for the linked list as an alternative way to construct a list. The linked list part is separated again into two sections, which are the usage of array for linked list and the usage of pointer for linked list.

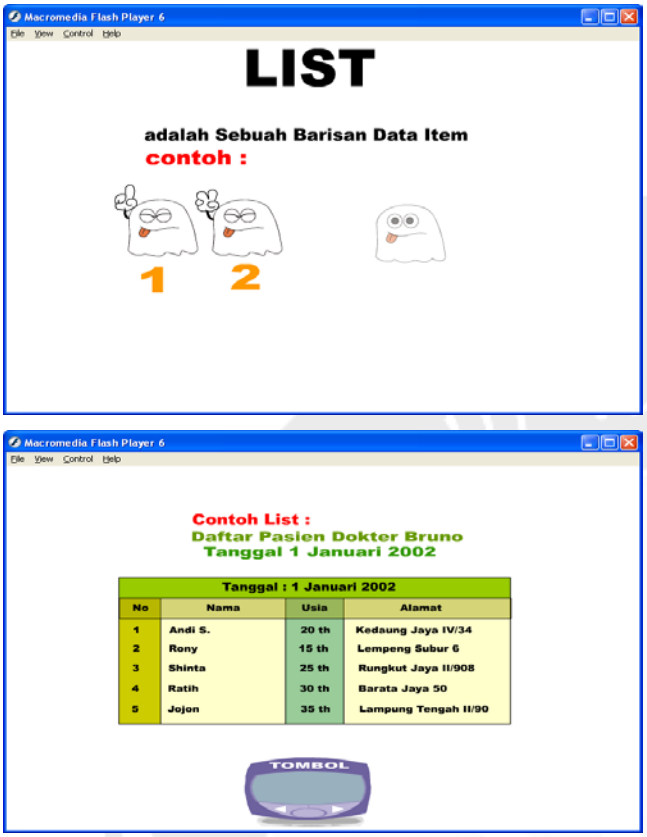


Figure 4. Introduction of List

On the introduction of list, the explanation is supported by some animation and example to describe it more clearly. The implementation of list definition and example is shown in Figure 4 above (the explanation is conducted in Bahasa Indonesia).

The implementation of sequential storage then follows the introduction of lists. The presentation can be started with illustration of the list implementation by implicit usage of ordering method. The example of the data put inside the storage is picture by some coins which would be delivered to the sequential storage, as shown in figure 5.

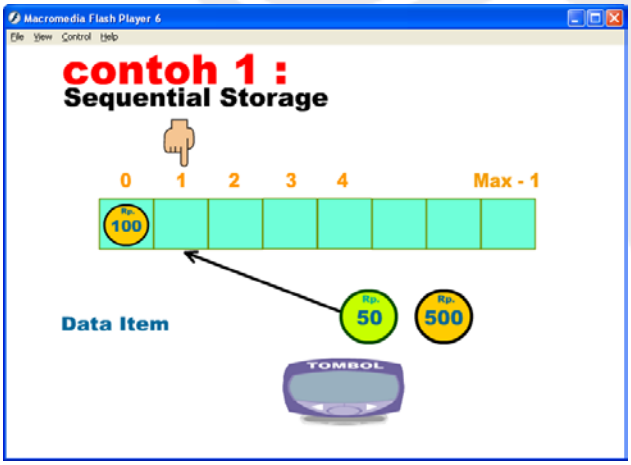


Figure 5. Illustration of sequential storage

The implementation of the sequential storage is covering the whole operations that should be provided for the list. Each operation would be explained by the means of process simulation and step by step debugging of the algorithm at the same time. Thus student can see the effect of every command line of the algorithm. The active command line is recognized by the used of red color for the line (figure 6).

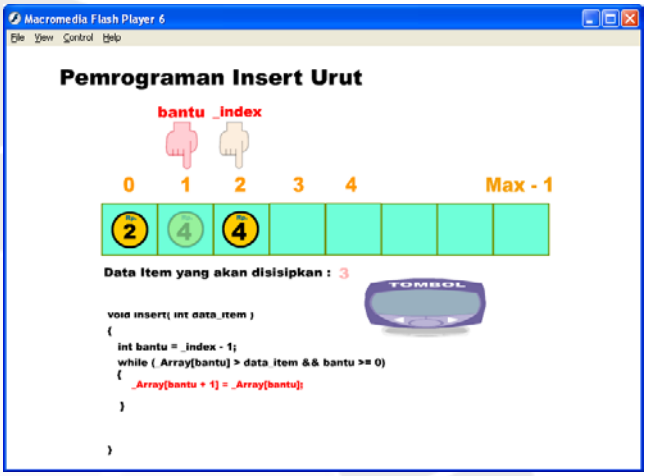


Figure 6. Insert process in sequential storage

The next section of the presentation is the linked list section. This part is divided into two parts. The first one is to explain the array based implementation of linked list. As the section previously, this part is also equipped with step by step execution of each line command and the simulation of the process behind it (figure 7).

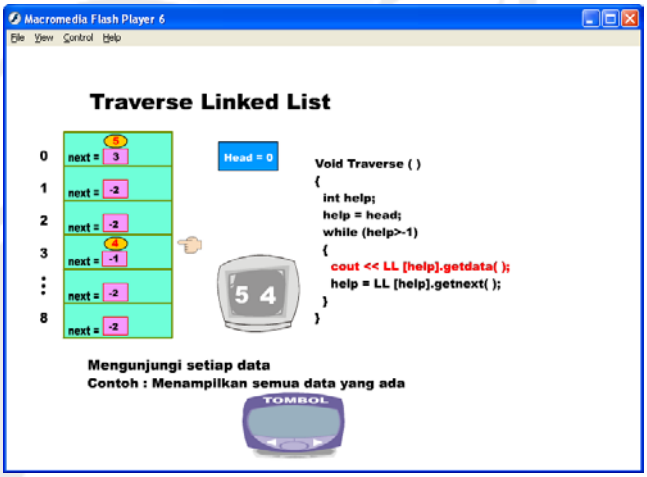


Figure 7. Traverse process in array based linked list

Implementation of pointer based linked list would be explained using a locomotive and wagons illustration and animation. Using this illustration student should be able to grab the understanding on how every element connected to each other and can be traverse starting from the locomotive as the head (figure 8).

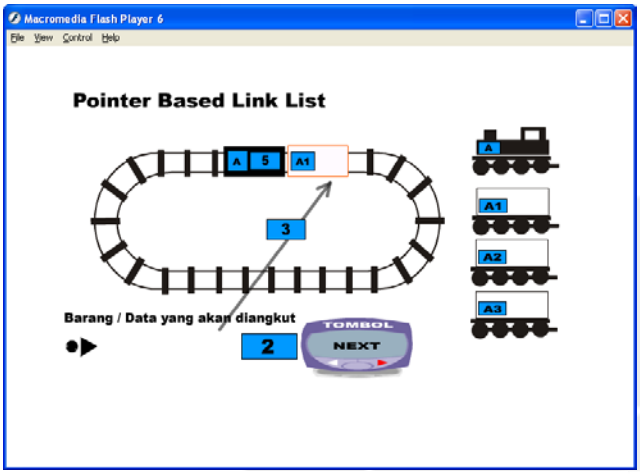


Figure 8. Illustration of pointer based linked list

Just like the other implementation of list operations, this part would also provide a simulation as an addition to step into command line method, as shown in figure 9 below.

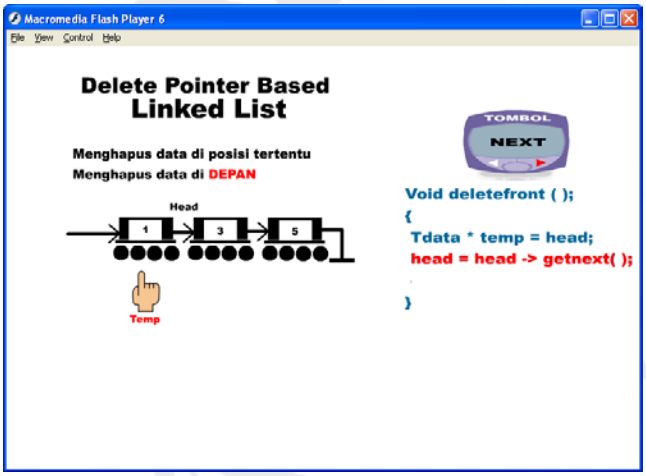


Figure 9. Delete element in pointer based linked list

6. CONCLUSION

Creating a computer aided learning for topic about list in data structure is a challenging project. The previous study has shown some obstacles which made the learning process of this subject harder. However, this barrier apparently can be overcome by the means of multimedia technology. Multimedia technology has proven itself to be adequate to bring the learning experience to the higher level in which the student has the chance to gain a better understanding through a better simulation and representation of the instruction material. And it offer a more interesting way of learning as well.

7. REFERENCES

[1] Nyhoff, L. and Leestma, S. 1992. Data Structures and Program Design in Pascal. Macmillan Publishing Company.

[2] Reddi, U.V. and Mishra, S. Ed. 2003. Educational Multimedia, A Handbook for Teacher-Developers. Commonwealth Educational Media Centre of Asia.

[3] Baldwin, L.V. and Down, K.S. 1981. Educational Technology in Engineering. National Academy Press.

Development Weightless Neural Network on Programmable Chips to Intelligent Mobile Robot

Siti Nurmaini
Department of Computer Engineering,
Faculty of Computer Science
University of Sriwijaya
siti_nurmaini@unsri.ac.id

Bambang Tutuko
Department of Computer Engineering,
Faculty of Computer Science
University of Sriwijaya
bambang_tutuko@unsri.ac.id

ABSTRACT
This paper presents an alternative hardware solution to be implemented on low cost microcontroller for mobile robot navigation. A RAM based weightless neural network (WNN) was considered as the heart of the controller caused by the advantage of its ease and simplicity implementation in cheaper hardware. The structure of the WNN was well appropriate to realize the experiment for intelligent mobile robots. The hardware implementation gives massive parallelism of neural networks and good recognition and low cost modification.

Keywords
Weightless neural network, environmental recognition, random access memory, intelligent mobile robot

1. INTRODUCTION
Mobile robots must recognize their environment in order to perform their tasks on the dynamic world. The recognition problem must be addressed to have robust performance because it cannot be ignored or avoided. Sensors are the only mean to identify the state of the environment. The problem of identifying the environmental state from sensor readings is often hard. Sensors usually have large amounts of noise in the readings they produce. Individual sensor readings are usually uninformative.

Autonomously recognizing in intelligent mobile robot is a prominent example of difficult realistic problems, and has long attracted the application of a wide range of powerful classification methods [18]. It is a very difficult problem needing a lot of computational power, and giving not so much accurate results in terms of robot pose estimation [19][20].

Mainstream artificial neural network (ANN) models are based on weighted-sum-and-threshold artificial neurons, as the pioneering *Threshold Logic Unit*, of McCullogh and Pitts [10]. The biological analogy behind this model lies on the mapping of the synaptic strength between the output produced and transmitted by the neuron's axon and the input of a post-synaptic neuron, into pseudo-continuous numerical weights [11]. Nevertheless generalizations of artificial weighted-sum-and-threshold neurons, such as Sigma-Pi units, do exist, this means that the dendritic tree, the mostly noticeable morphological structure of the neuron cell, is not being taken into account in mainstream ANN paradigms [11].

Weightless neural networks (WNNs) are based on networks of Random Access Memory (RAM) nodes. WNNs are a variant of artificial neural networks that are trained to recognize a pattern based on lookup tables that store neuronal functions. They do not have multiplicative weights between nodes, hence the name [6].

The use of RAM nodes in pattern recognition problems is dating 50 years by the work of Bledsoe and Browning [1]. These networks are typically used in pattern recognition applications because of their small size and computational requirements [4]-[5]. Some years later, Aleksander introduced Stored Logic Adaptive Microcircuit (SLAM) and n-tuple RAM nodes as basic components for an adaptive learning network [12]. With the availability of integrated circuit memories in the late 70s, the WiSARD (Wilkes, Stonham and Aleksander Recognition Device) was the first artificial neural network machine to be patented and produced commercially [13][14]. Other WNN models followed, such as PLNs [15], GSNs [16] and GRAMs [17].

Many researcher using this technique in mobile robot application indicates this technique maybe successful such as, [2],[3],[7],[8],[9],[20]. This paper demonstrates the potential technique of WNN in embedded mobile robot for recognizing and classifies the environment. In the paragraphs that follow, the WNN structure and application in mobile robot will be presented. Results of experiments that measure classification of environment will also be given.

2. WEIGHTLESS NEURAL NETWORK

2.1 Definition
The main component of the WNN is the ensemble of class discriminator (Figure 1). A discriminator consists of a series of address spaces similar to Random Access Memory (RAM) components that are attached to each of the n-tuples, where the value of the n-tuple is employed as the address of the RAM location, being incremented when the network is trained. The neuron inputs are connected in a random sequence to the feature vector, each neuron is a binary pattern recognition device. Each discriminator consists of M RAM-like neurons (weightless neurons) with n address lines, 2^n storage locations (sites) and 1-bit word length. Each RAM randomly samples n bit of the input pattern. Each pattern must be sampled by at least one RAM.

For an input vector of size K, the number of necessary neurons J of connectivity N that should be used to cover all inputs of the input vector should satisfy: $J \times N > K$. This neuron group is called a discriminator and its response is produced by connecting an adder that sums the neuron outputs, counting the number of active neurons (neurons outputting "1") in the group [6]. This response vector can be regarded as a feature vector that measures the similarity of an input pattern to all classes. In the WNN, a Winner-Takes-All-Block can be attached to the adder outputs to choose the discriminator containing the greater number of active neurons, pointing to the winning classes. Each pattern will produce a feature vector that describes its similarity to all classes.

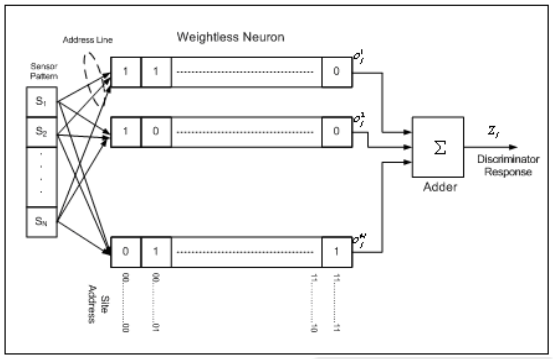


Figure 1. RAM Discriminator

2.2 Neuron Architecture

To implement the WNNs, the 8 bit data for each sensor was used. The WNNs have neuron structure as shows in Figure 2. The neurons are connected in groups (discriminators) that correspond to one of the possible classes of commands the neural network can choose. The groups are connected to an output adder (o1, o2, …,on) that counts the number of active neurons in the group. For each class of input patterns one discriminator is needed, which is trained solely on the input data of its own class.

The different between two binary patterns is the number of bit positions in which they differ and this gives us a rough idea of how similar the two patterns are the measure of similarity to correlate the inputs to the neuron and a stored value. Winner take all decision is chose the active neuron.

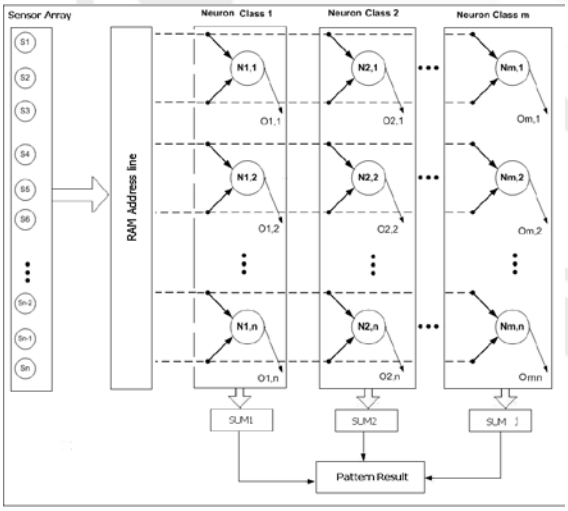


Figure 2. Architecture of neurons

The input address vector is presented to the network. The desired output of every cell in is the *same* as the desired output. If the desired output at the input layer is **1**, the addressed location is incremented by one. If the desired output is **0**, the location is decremented by one. The learning algorithm then involves calculating the address vectors for the next layer, moving towards the output

During the recalling phase, again, the content of each addressed location, starting from the input layer, is interpreted as U (undefined), 0, or 1 as the counter value is zero, negative, or positive respectively. During recall, we can clearly see that the output of each cell might propagate forward an undefined output. In the output cell, we determine, for each U location addressed, the nearest address up to a Hamming distance measure of *d* to a defined location's address (either 0 or 1). We continue to run through the training set in this way until all U's are replaced by 1's or 0's in the output layer. The size of *d* should be less than half the cell size. We have found that the overall classification result improves significantly.

3. ENVIRONMENTAL RECOGNITION

We considered the common target that there exist in real environment of mobile robot applications such as plane, edge, corner with angle 90 degree, acute corner with angle 60 degree. Length of plane is about 45 cm and other objects are of similar size. These objects at four distances: 10, 20, 30, 40 cm. Also angle between the head of mobile robot and these objects is assumed to be -30, -20, -10, 0, 10, 20, 30 degree. Fig. 2, show the environmental classification at these nine classes environment.

The WNNs is designed to identify the current environment by recognizing typical patterns. To implement the network, 8 bit data from eight ultrasonic sensors is used to determine the direction of the obstacle. The combination of them appearance in the seven directions makes up different input pattern such as, front, right front, left front, right front side, right back side left front side and left back side. The winner-takes-all decision chooses that has more active neurons and encodes it.

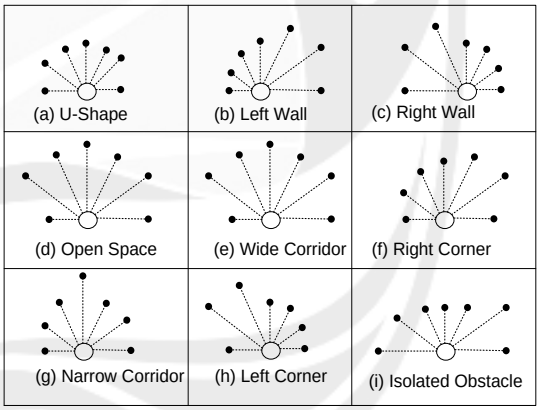


Figure 3. Environmental pattern

The patterns with a single far, medium, or near obstacle to train the neural network is used in this research. At a time the obstacle is placed in different directions and in difference distance. The WNNs then is taught that the obstacle at left side, right side or forward. Using this technique, the value distinguishing distance an obstacle has to be obtained first. In this experiment, the variable sensors are single byte that holds the sensor readings. In the evaluation phase, the WNNs by generating all the possible input combination. The number of possible combination is 2^8 = 256 combination. Then, each output for all input possibilities was written to a lookup table, representing the neuron combination.

The calculation of this value is based on the distance from an obstacle to the robot, Using on this calculation, the threshold values for distance of the robot are 00011110 (30 cm) indicates the obstacle is far, 00010100(20 cm) the obstacle is medium, and

00001010 (10 cm) the obstacle is near and 01001011 (75 cm) no obstacle is detected.

4. EXPERIMENTAL RESULT

The designed WNN was evaluated in several experiments involeved with mobile robots navigation task. A 2.0 m x 5.0 m area containing walls corridor was created as mobile robots working domain. A 112 bits WNN architecture was chosen with 7 classes, 2 neurons per class and 8 bits per neuron. Seven classes indicate seven direction of the obstacle in the environment. The binary pattern or combination of 1 and 0s in the 168 bits WNN architecture defines the pattern of environment. The contents of the neuron were initially either randomized or set to zeros and the effects of the training on such WNN were observed and the environmental recognition effects on training were investigated.

Table. 1 Critical class of environmental recognition

Actual Place	Distance (cm)	Refe-Rence (hex)	Result (hex)
Convex (90^o)	20	12h	12h
	30	0bh	0bh
	40	05h	0dh
Concave (270^o)	10	06h	00h
	20	0eh	00h
	30	00h	00h
Plane (180^o)	10	03h	07h
	20	0ch	0ch
	30	15h	15h
	40	1eh	1eh
Left-corner (180^o)	10	03h	07h
	20	0ch	0ch
	30	15h	15h
	40	1eh	1eh
Right-corner (180^o)	10	03h	07h
	20	0ch	0ch
	30	15h	15h
	40	1eh	1eh
Corridor (0^o)	10	03h	07h
	20	0ch	0ch
	30	15h	15h
	40	1eh	1eh
U-shape (180^o)	10	03h	
	20	0ch	
	30	15h	
	40	1eh	

Experiment is conducted to demonstrate the ability of a mobile robot to react to various unknown environment. The result is based on the environment classification. Table 1 has shown using 10 experiment data, WNNs approach has achieve 94 % classification. How ever the poorest result was if the robot closes the object, where the scanning sensory sector of the robot was quite high and some noise has still interfered in echo signal.

The performance of WNNs also can be seen from the pattern rejected besides the success in recognizing patterns. The better quality of WNNs determined, also by the greater of rejected patterns from the recognition. Experiments is conducted as much as 10 times and taken the average value of a recognizable pattern number of environmental patterns expressed in the input changed from 4, ,and 9 patterns, while the neurons are used changed from at least 3 neurons to 10 neurons. Percentage of success for the rejected pattern can be seen in the Figure 4.

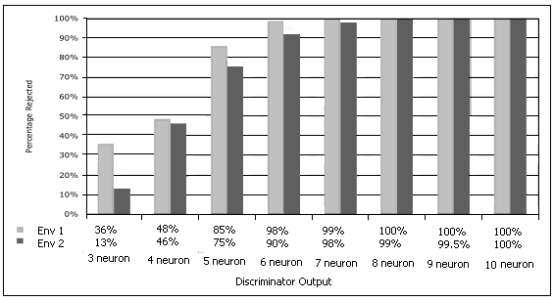


Figure 4. Pattern rejected for 4 patterns and 9 patterns of environment

Greater number of neurons gives better environment pattern rejected, up to 99% and 100%. But, the usage of more neurons will increase memory usage in detection process. In the future work the number of optimal neuron will be investigated for better generalization without memory saturation.

5. CONCLUSION

We have demonstrated the potential technique of WNNs in embedded mobile robot for recognizing and classify the environment and its successful application in sensory-motor navigation task for mobile robots. This is an alternative and unique hardware implementation of the neural network inside low cost microcontroller to allow fast response and make it available in a single chip. By designing appropriate architecture the same basic of the WNN could be used to realize low cost hardware implementation in real time mobile robot.

6. ACKNOWLEDMENTS

This work was supported by Faculty of Computer Science, Sriwijaya University, Indonesia

7. REFERENCES

[1] Bledsoe, W.W., and Browning, I. 1959. Pattern Recognition and Reading by Machine, Proceedings of the Eastern Joint Computer Conference, Boston, pp. 225-232.

[2] Mitchell, R.J., Keating, D.A., and Kambhampati, C.1994. Neural network controller for mobile robot insect. Internal report, Department of Cybemetics, University of Reading.

[3] Zhou, Y., Wilkins, D., Cook, R.P.1998. Neural network for a fire-fighting robot. University of Mississippi..

[4] I. Aleksander, W. V. Thomas, P. A. Bowden, "Wisard: A radical step forward in image recognition." Sensor Review, pp 120-124, July 1984

[5] J, Austin, "A Review of RAM-Based Neural Networks". Proceedings of the Fourth International Conference on MICRONEURO94, pp. 58-66, 1994.

[6]. L. Teresa et al, "Weightless Neural Models: A Review of Current and Past Works", Neural Computer Surveys v. 2,pp. 41-61, 1999

[7]. Simoes, E. D. V., Uebel, L. F., and Barone, D. A. C., Hardware Implementation of RAM Neural Networks. In Pattern Recognition Letters, n. 17, pp. 421-429, 1996

[8] Botelho, S. C., Simoes, E. D. V., Uebel, L. F., and Barone, D. A. C., High Speed Neural Control for Robot Navigation. Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, Beijing, China, pp. 421-429, 1996.

[9] Q. Yao, D. Beetner, D.C. Wunsch, and B. Osterloh., A RAM-Based Neural Network for Collision Avoidance in a Mobile Robot, IEEE, 2003.

[10] McCulloch, W. and Pitts, W., A logical calculus of the ideas immanent in nervous activity, Bulletin of Mathematical Biophysics, 7, pp. 115-133, 1943.

[11] I. Aleksander, "Weightless Neural Network", proceedings, European Symposium on Artificial Neural Networks - Advances in Computational Intelligence and Learning, Belgium, 2009.

[12] I. Aleksander, Ideal neurons for neural computers, in: Parallel Processing in Neural Systems and Computers, North-Holland, Amsterdam, pp. 225-228, 1990.

[13] I. Aleksander and T.Stonham, Guide to pattern recognition using random-access-memories Computer and Digital Techniques, 2, pp. 29-40, 1979.

[14] I. Aleksander, W. Thomas, and P. Bowden, WISARD, a radical new step forward in image recognition, Sensor Rev., 4(3), pp. 120-124, 1984.

[15] I. Aleksander and W.W. Kan, A Probabilistic Logic Neuron Network for Associative Learning, IEEE Proceedings of the First Int. Conf. on Neural Networks, pp. 541-548, 1987.

[16] R. G. Bowmaker and G. G. Coghill, Improved recognition capabilities for goal seeking neuron, IEE Electronics Letters, 28, pp. 220-221, 1992.

[17] I. Aleksander, Ideal neurons for neural computers, in: Parallel Processing in Neural Systems and Computers, North-Holland, Amsterdam, pp. 225-228, 1990.

[18] Borenstein, J., Everett, H.R., Feng, L.: Navigating Mobile Robots: Systems and Techniques. A. K. Peters, Ltd., Natick, MA, USA.,1996.

[19] Thrun, S., Burgard, W., Fox, D.: Probabilistic Robotics (Intelligent Robotics and Autonomous Agents). The MIT Press, 2005.

[20] Nurmaini, S., Zaiton, S., Norhayati, D. Siti Nurmaini, Siti Zaiton, Dayang Norhayati. RAM Network based type2 Fuzzy-Neural Controller for Navigation of Mobile Robot", Proceedings of International Conference Control And Mectronic. 2009., Malaca, June 2009. pp. 392-397.

If-Statement Modification for Single Path Transformation: Case Study on Bubble Sort and Selection Sort Algorithm

Rahmadi Trimananda
Universitas Pelita Harapan
Jl. M.H. Thamrin No.2
Lippo Karawaci, Tangerang 15811
Phone. 62-21-5460901 ext. 2300
rahmadi.trimananda@staff.uph.edu

ABSTRACT
Timing analysis, namely worst-case execution time (WCET) analysis, is a focal research theme in real-time systems feasibility assessment. Despite its high importance, it exhibits high complexity of analysis as the software gets more complex. High complexity and unpredictability in timing analysis arise as the program code contains a lot of possible paths, which are dependent on the input values. With the single execution path paradigm, timing analysis can be made simpler. Throughout this paper, Bubble Sort and Selection Sort algorithms are presented, modified in single path approach, and tested. The if-statements are the targets of this modification. The experiment shows that it is possible to generate single path codes without having to modify the low level code and the results signify improvements on the temporal predictability.

Keywords
real-time system, algorithm, flow control, genetic algorithm, static timing analysis

1. INTRODUCTION
The current development of computer technology actually conceals the complexity of the underlying systems. Complex systems, e.g. aircraft navigation system and robotic control system, actually consist of sophisticated management and handling of real-time actions. As computers become increasingly more powerful, people try to create more complicated applications that need real-time scheduling and executions.

For handling this real-time issue, real-time systems emerge to help schedule the actions that needed to be accomplished at the right moments. Therefore, predicting and determining the execution times of each action are very important to make the correct scheduling that meets all deadlines. This holds, especially for hard real-time systems, which are not tolerant to exceeding deadlines. In this case, worst-case execution time (WCET) measurement plays an important role in timing analysis as it presents the worst possible execution time of a program code.

Unfortunately, WCET analysis becomes more complicated as the program gets more complex. The execution times of the program vary greatly, which is due to a variety of techniques applied to make computer hardware run more efficiently and many possible paths existing in a program code.

Therefore, the single path execution paradigm [7] can be applied to algorithms to reduce the fluctuation of execution times, which then will ease the complexity of WCET analysis. However, if the algorithm is modified into a single path algorithm, the execution time can be longer because it offers one path for all kinds of input combination, including simple combinations that would have taken less time to execute. Therefore, choosing the correct methods for applying the single path execution paradigm is important if we want to realize a feasible real-time system and, yet, analyzable.

The experiment explained in this paper is intended to observe this single path paradigm. Two sorting algorithms, i.e. Bubble Sort and Selection Sort, are taken as objects of the experiment. They are modified into single path algorithms to be less data dependent and more temporally predictable.

The following sections will discuss about this single path experiment in greater details. Section 2 is dedicated for discussing the basic concepts of timing analysis, automated testing, and single path paradigm in relation to the sorting algorithms. Section 3 presents the experimental setup and results, along with the evaluation. Finally, section 4 completes the paper with summary and conclusions.

2. TIMING ANALYSIS IN REAL TIME SYSTEMS

2.1 Timing Analysis
In real-time systems, timing analysis is most likely to be the central point of attention. It is usually performed both in static and dynamic timing analysis, and the main objective is to find the worst-case execution time (WCET). In real-time system software, WCET gives a prediction of the longest execution time. This is really important because this execution time prediction will assess the feasibility of the corresponding real-time system software [15].

2.2 Static Timing Analysis and Dynamic Timing Analysis
WCET is usually conducted in two ways: static and dynamic timing analysis [12]. Static timing analysis is accomplished by analyzing the program code thoroughly, while dynamic timing analysis is carried out by running the compiled code. Normally, static timing analysis provides a boundary value for dynamic timing analysis the static analysis indicates an execution time that is usually longer than the one obtained by actually running the program.

Static analysis gives a possibility to measure the execution time of a program, without having to run it on a specific platform [12]. The execution time is determined by reading the program, analyzing its

parts, and taking into account different possible input combinations with respect to their different effects on the program execution paths. It becomes increasingly more difficult as the program gets bigger and more complex.

On the other hand, dynamic timing analysis is a simple way of measuring program performance, i.e. run the program and measure the execution time. However, the results usually vary and, thus, it is difficult to justify the accuracy of the measurement. Moreover, it is really dependent on the platform, on which the program is running. Nonetheless, it gives an idea of the actual execution time of the running program on a certain platform. If the measurement is done carefully and the results are interpreted thoughtfully by taking all possible aspects into account, the aptness of the program for real-time systems can be assessed properly.

2.3 Genetic Algorithm in Automated Software Testing for Timing Analysis

The difficulty in performing dynamic timing analysis lies on producing good test sets that are powerful enough to explore all possibilities of the program execution. Manual method is tedious and, thus, not preferred in attempting for an accurate result. To address this problem, one may consider applying automated software testing to ease the burden, e.g. random testing, evolutionary testing, etc. Aside from the random testing, which is more likely to be not deterministic and inconclusive due to its indefinite behavior, evolutionary testing based on genetic algorithm can be a good choice for automating the testing procedure and, yet, coming up with accurate testing results [6].

In general, evolutionary testing exploits the potential of genetic algorithm, which involves several important parameters to choose and a number of steps to carry out. Imitating the process of evolution that is believed to have been happening in the nature, a genetic algorithm performs evolution on a population of bit strings that are usually called chromosomes. The steps carried out on the chromosomes consist of initialization, selection, recombination, and mutation. When implementing genetic algorithm, one should pay attention to some important parameters that determine the success of this method: population size, number of generations, crossover rate for recombination, fitness function, and mutation rate.

2.4 Single Execution Path to Ease Timing Analysis

As the program code becomes more complex, WCET analysis also becomes more difficult. The complexity of WCET analysis comes from both software and hardware sides [3]. A lot of possible sequences of actions in a program and the execution times needed for each one of them contribute to the complexity from the software side. This condition is made even worse by the presence of advanced hardware features, e.g. branch prediction and cache memory. They can also give bizarre effects due to their indigenous dynamic behavior, unless if they are made static [4] [5].

To simplify the complexity of WCET analysis, single execution path transformation [2] can be considered as a good solution. It removes the above mentioned problems of software sides as it makes the program execute only a single possibility of path; thus, it only has one possible execution time. However, the resulted program often becomes far less efficient than the original one if the transformation is done improperly. For most cases, the program execution time becomes longer because it always takes one sequence of actions, even for some simple cases that do not need certain actions in the sequence.

The experiment explained in this paper is intended to introduce one way of transforming if-statements that absolutely generate more than one execution path. For this purpose, two simple sorting algorithms with if-statements, i.e. Bubble Sort and Selection Sort, are modified to get rid of multiple execution paths. Those if-statements make the two algorithms still somewhat dependent on the input values. This condition summons problems, when proper timing analysis needs to be done.

There is some research done on various sorting algorithms, in terms of their appropriateness for real-time applications [13] [14]. Those previous experiments show that the algorithms that are less dependent on input data perform more stably. In other words, they have more predictable execution time.

Some analyses and observations are also carried out to formulate good methods on algorithm modifications towards more predictable execution times [1] [2] [7] [8]. In [1] [2] [7], the algorithms are modified by transforming conditional statements into predicated instructions, i.e. conditional move. However, this involved changing the lower level assembly code as there was no compiler that could generate such instructions. Nevertheless, those experiments yielded quite promising results on single path paradigm as it has succeeded in making execution times more predictable.

The following sections explain an experiment on Bubble Sort and Selection Sort algorithms. It shows a possibility of transforming input dependent algorithms into temporally predictable codes without even having to crack the low level code.

3. EXPERIMENTAL WORK

3.1 Bubble Sort and Selection Sort Algorithms and Transformations

As explained in Sect. 2.4., both Bubble Sort and Selection Sort algorithms have if-statements that reduce the predictability of the timing analysis. This experiment attempts to remove the time unpredictability caused by the if-statement, so the final goal is to remove the if-statements at all. To perform this experiment, both algorithms are modified twice. In the first modification, the if-statements are changed into if-else-statements, so it can be assumed that one of the two existing paths (if-path or else-path) in the branching statements is always taken. In this case, only the if-path contains really essential instructions; the else-path is made as a dummy path for those input elements that do not exercise the if-path. By doing this, the program will always take the longest path for any combinations of input values.

In the second modification, aside from what is usually done for single path transformation [1] [2] [7], which involves the use of conditional move instruction by changing the assembly code of the program, this experiment applies a different transformation on the if-statements. In sorting algorithms, an if-statement usually compares two elements to decide whether to swap them or keep them on their original positions. The condition for taking decisions

yields a Boolean value in binary: 0 or 1. If this condition is taken out and transformed into a calculation for generating a swapping index of the two elements, the if-statement can surely be eliminated. Thus, the code transforms into a single path code. The following excerpts show the second modification of the two algorithms. The modified parts are printed in bold.

Algorithm 1: Bubble Sort

```
procedure BubbleSort( A : list of sortable items )
  for each i in 0 to length(A) do:
    for each j in 0 to i do:
      if (A[j] > A[j+1])      ==> int swp_ind = j + (A[i] > A[i+1])
        swap(A[j], A[j+1]) ==> swap(A[i], A[swp_ind])
    end for
  end for
end procedure
```

Algorithm 2: Selection Sort

```
procedure SelectionSort( A : list of sortable items )
  for each i in 0 to length(A) do:
    min = i;
    for each j in i+1 to length(A) do:
      if (A[j] < A[min]) ==> index[2] = j,min
        min = j;          ==> swp_ind = 1 - (A[j] < A[min])
                          ==> min = index[swp_ind]
    end for
    swap( A[ i ], A[min] )
  end for
end procedure
```

Figure 1. Bubble Sort and Selection Sort algorithms

Therefore, the algorithms are tested in three different ways of implementation. In this case, the three versions of Bubble Sort algorithms are ordinary Bubble Sort, if-else-statement modified Bubble Sort, and Bubble Sort without if-statement, whereas the Selection Sort also has those three versions.

3.2 Experimental Setup

The algorithms are implemented in C and compiled by the GNU GCC compiler without allowing any optimizations to take place. The genetic algorithm is implemented according to the methods and genetic algorithm discussed in [6] and [9]. This genetic algorithm is performed in 200 generations with a population size of 50, crossover rate of 0.5 and mutation rate of 0.1. The intention is to produce test sets to feed the sorting algorithms, i.e. three versions of Bubble Sort and three versions of Selection Sort.

The programs are executed under RTAI-Knoppix Linux environment on an Intel Core Duo machine running at 1.86 GHz. The RTAI used is from version 3.4 [10] [11] running on a Linux Kernel of version 2.6.17.11. In the program code, a subprogram is added to execute the program in a hard real-time environment on the kernel space to give even more assurance to the real-time execution. Additionally, the execution is given the highest priority among any other processes.

3.3 Results and Evaluation

The experiment is conducted in different numbers of array elements: 5, 10, 50 and 100 elements. Best case execution times (BCET), average case execution times (ACET), and worst case execution times (WCET) are taken for comparing the performances of the algorithms. Sorting operations for 5, 10, and 50 elements show that the six implementations give approximately the same level of performance. However, interesting results, as shown in Table 1. and Figure1, appear for the sorting operations with 100 elements.

Table 1. Execution times (in ns) of sorting operations on 100 array elements

Algorithms	BCET	ACET	WCET
BubbleSort1	72914.000	2890242.000	5290065.000
BubbleSort2	106439.000	4734943.540	7338373.000
BubbleSort3	104762.000	466467.660	1399621.000
SelectionSort1	44419.000	2512579.980	3784006.000
SelectionSort2	59505.000	2626309.800	4641379.000
SelectionSort3	108953.000	201344.220	688077.000

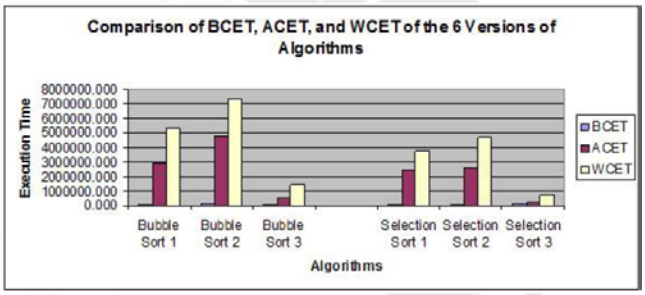


Figure 2. Comparison of the execution times in a bar chart

The comparison of execution times of the ordinary versions of both Bubble Sort and Selection Sort show relatively big differences between best, average, and worst case execution times. These differences in execution times, usually called time jitter, are most likely to be inflicted by the advanced hardware features (Sect. 2.4.). This condition also occurs in the execution times of the first modifications of both algorithms, which deploy if-else-statements. On the contrary, the second modifications of both Bubble Sort and Selection Sort show relatively stable execution times compared to the first two versions.

Conclusively, the attempt to completely omit the if-statements in the two algorithms, as it is done in the second modification, shows its predominant effect. It we assume that the hazardous effect on time analysis from if-statements is completely nullified, the source of time differences that are still evident might come from the other uncontrolled hardware features, e.g. caching. This indicates that the attempts to improve program execution time in software have to be also supported by hardware enhancement. On the other hand, the first modification that uses if-else statements seems to be not suitable and, thus, not preferable for single-path transformation as it still even suffers from branch ludicrous effects on execution times.

4. SUMMARY AND CONCLUSIONS

As it is difficult to assess the quality of timing analysis due to data dependent execution times imposed by different execution paths that can be taken by a program, having a single execution path will

certainly resolve the jeopardy. With only one execution path, timing analysis can be far simpler and easier to do. By applying a correct transformation the algorithm, one can modify the code to have a single execution path without even touching the low level code.

The experiment conducted on the ordinary versions of Bubble Sort and Selection Sort, together with two other modifications for each algorithm has shown that the single path transformation has been made possible to eliminate the serious effects of if-statement on execution times, due to the presence of branch predictions. Some time differences might still appear because of other possibilities, e.g. caching. Therefore, one should bear in mind that software efficiency techniques must also be accompanied by hardware improvement. Nonetheless, the alternative if-statement modification for single path execution has shown a significant improvement on the algorithms.

There are many things that can still be observed in the future. Different algorithms, apart from the two algorithms presented above, will require different approaches for single path transformation. Therefore, a structured methodology needed for this single path transformation on various algorithms will still be an interesting topic for future work.

5. ACKNOWLEDGMENTS

I would like to thank Gerhard Gross from the Software Engineering Department of TU Delft, Netherlands. His Real-time System course really gave me insights in finishing this paper.

6. REFERENCES

[1] Peter P. Puschner. Transforming Execution-Time Boundable Code into Temporally Predictable Code. DIPES ’02: Proceedings of the IFIP 17th World Computer Congress - TC10 Stream on Distributed and Parallel Embedded Systems. Kluwer, B.V., 2002.

[2] Peter Puschner and Alan Burns, Writing Temporally Predictable Code. Proc. 7th IEEE International Workshop on Object-Oriented Real-Time Dependable Systems. p.85–91., Jan. 2002.

[3] Peter Puschner, Is Worst-Case Execution-Time Analysis a Non-Problem? – Towards New Software and Hardware Architectures. Proc. 2nd Euromicro International Workshop on WCET Analysis. Department of Computer Science, University of York, 2002.

[4] Puaut, I., Cache Analysis vs Static Cache Locking for Schedulability Analysis in Multitasking Real-Time Systems. Proc. of the 2nd International Workshop on worstcase execution time analysis, in conjunction with the 14th Euromicro Conference on Real-Time Systems, June 2002.

[5] Claire Burguire and Christine Rochange and Pascal Sainrat, A Case for Static Branch Prediction in Real-Time Systems. Real-Time Computing Systems and Applications, International Workshop on. p.33-38. IEEE Computer Society, 2005.

[6] H.G. Gross, An Evaluation of Dynamic, Optimisation-based Execution Time Analysis. Proceedings of Intl. Conf. on Information Technology: Prospects and Challenges in the 21st Century (ITPC-2003). Nepal Engineering College, 23-26 May 2003.

[7] Peter Puschner, Experiments with WCET-Oriented Programming and the Single-Path Architecture. WORDS ’05: Proceedings of the 10th IEEE International Workshop on Object-Oriented Real-Time Dependable Systems. p.205–210. IEEE Computer Society, 2005.

[8] J. R. Allen and Ken Kennedy and Carrie Porterfield and Joe Warren, Conversion of control dependence to data dependence. POPL ’83: Proceedings of the 10th ACM SIGACT-SIGPLAN symposium on Principles of programming languages. ACM, 1983.

[9] Zbigniew Michalewicz, Genetic Algorithms + Data Structures = Evolution Programs - 3rd Ed. Springer, 1998.

[10] RTAI Team, RTAI - the RealTime Application Interface for Linux from DIAPM. http://www.rtai.org, 2008.

[11] Giovanni Racciu and Paolo Mantegazza, RTAI 3.4 User Manual rev 0.3. http://www.rtai.org, 2006.

[12] Daniel S and Andreas Ermedahl and Jan Gustafsson and Bjrn Lisper, Static timing analysis of real-time operating system code. In 1st International Symposium on Leveraging Applications of Formal Methods (ISOLA04), 2004.

[13] Dietmar Mittermair and Peter Puschner, Which Sorting Algorithms to Choose for Hard Real-Time Applications. Proc. 9th Euromicro Workshop on Real-Time Systems. p.250–257, Jun. 1997.

[14] Peter Puschner and Alan Burns, Time-Constrained Sorting - A Comparison of Different Algorithms. Real-Time Systems, Euromicro Conference on. p.78. IEEE Computer Society., 1998.

[15] Peter Puschner and Alan Burns, A Review of Worst-Case Execution-Time Analysis. Journal of Real-Time Systems. p.115–128, May 2000.

Implementation of Particle Swarm Optimization Method in K-Harmonic Means Method for Data Clustering

Ahmad Saikhu
Informatics Department,
Faculty of Information Technology
Institut Teknologi Sepuluh Nopember
(62-31-5999214)
saikhu@its-sby.edu

Yoke Okta
Informatics Department,
Faculty of Information Technology
Institut Teknologi Sepuluh Nopember
(62-31-5999214)
Yoke_okta@cs.its.ac.id

ABSTRACT

Clustering is a method for partitioning a set of objects into homogeneous groups (clusters) based on a specified set of variables. Goals of this method is objects within a cluster are similar and dissimilar with the objects in other clusters. K-Harmonic Means (KHM) is a clustering algorithm that can solve problems on the cluster center initialization of K-Means algorithm, but KHM still can not overcome local optima problem. Particle Swarm Optimization (PSO) is a stochastic algorithm that can used to find optimal solution to a numerical problem, but PSO has a problem at the convergence speed.

To overcome these problems, there is Particle Swarm Optimization K-Harmonic Means (PSOKHM) algorithm which is a combination of KHM and PSO algorithm. In this final project, PSOKHM algorithm used to perform data clustering, and KHM and PSO algorithm as a comparison for evaluation of the cluster-based objective function value, F-Measure, and the running time. Trials conducted with 3 scenarios of 5 different data sets. From the result of the test obtained that, when viewed from the objective function and F-Measure value, PSOKHM able to give better. Meanwhile, if viewed from the running time, PSOKHM surpasses PSO but it is not better than KHM.

Keywords

Data Clustering, K-Harmonic Means, Particle Swarm Optimization

1. INTRODUCTION

Clustering is the process of grouping data objects into different classes, called clusters so that objects that are on the same cluster more similar and different from objects in other clusters. K-Means (KM) is one of the most popular algorithms used for clustering because of the feasibility and efficiency when dealing with a lot of data. Although the algorithm is easily implemented and can work quickly in many situations, KM algorithm has several weaknesses, including the results of the cluster is sensitive to the initial cluster centers and the results may lead to local optima [2].

To overcome the problems that occur in the initial cluster centers, Zhang, Hsu, and Dayal (1999.2000) [8] proposed a new algorithm called *K-Harmonic Means* (KHM) , and then modified by Hammerly and Elkan (2002). The purpose of this algorithm is to minimize the harmonic average of all points on the entire data set to the cluster center. Although KHM algorithm can solve the initial problem, KHM still can not overcome local optima [2]. *Particle Swarm Optimization* (PSO) is a stochastic algorithm designed by Kennedy and Eberhart (1995), which was inspired by the behavior of a flock of birds [4].

In this paper, the authors explore how the PSO algorithm helps KHM algorithm to move away from local optima. By using these two algorithms, a hybrid data clustering algorithm called Particle Swarm Optimization K-Harmonic Means (PSOKHM) was introduced. Based on test results on several data sets, obtained that the results of the PSOKHM algorithm is better than KHM and PSO. PSOKHM algorithm is not only overcome the local optima problem in KHM algorithm, but also increase the speed of convergence of PSO algorithm [2].

This paper can be divided as follows: Section 2 introduces the KHM clustering algorithm. In section 3 explains how the PSO algorithm is used in the clustering process. Section 4 describes hybrid algorithm PSOKHM. Section 5 contains the implementations and test results of the 5 data sets, namely Iris, Glass, Cancer, CMC, and Wine. Then the last, section 6 contains conclusions.

2. K-HARMONIC MEANS

K-Harmonic means is one of center based clustering method introduced by Zhang in 1999 which later developed by Hammerly and Elkan in 2002. The purpose of this algorithm is to minimize the harmonic average of all points on the entire data set to the cluster center. In the K-Means, each data point added only to one centroid, which means that each data point only has relevance to the centroid where the data is inserted. In the local area with high density of data points to centroid, centroid has the possibility can not move from one data point despite the fact that there are two nearby centroid. This second centroid may have worse local solution, but the global effect from replacement of the centroid might be useful for the clustering process to obtain better results [8].

In the KHM algorithm, the distance from each data point to all centroid is computed. The harmonic average is sensitive to the fact that there is two or more centroid is located near a data point. This algorithm is naturally exchange one or more centroid to an area where there is a data point that does not have a nearby centroid. So the better the results of cluster, the value of objective function will be smaller [8]. The following notation is used to formulate the KHM algorithm [2]:

$X = \{x_1,...,x_n\}$: the data to be clustered
$C = \{c_1,...,c_k\}$: the set of cluster centers
$m(c_j|x_i)$: the membership function defining the proportion of data point x_i that belongs to center c_j.
$w(x_i)$: the weighting function defining how much influence data point x_i has in re-computing the center parameters in the next iteration. KHM modul can be seen in Figure 1 and 2.

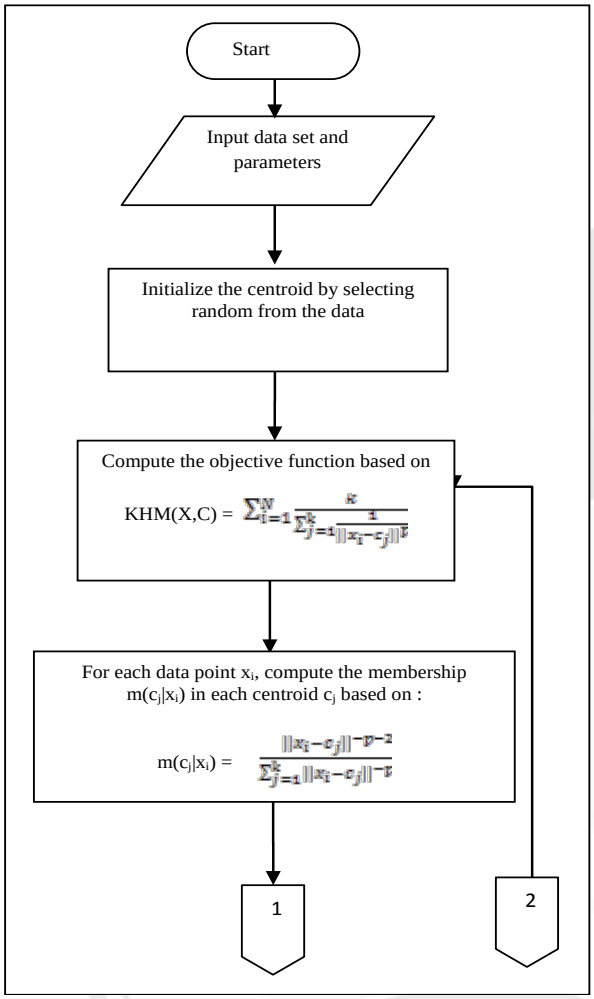


Figure 1. Flowchart KHM module sect.1

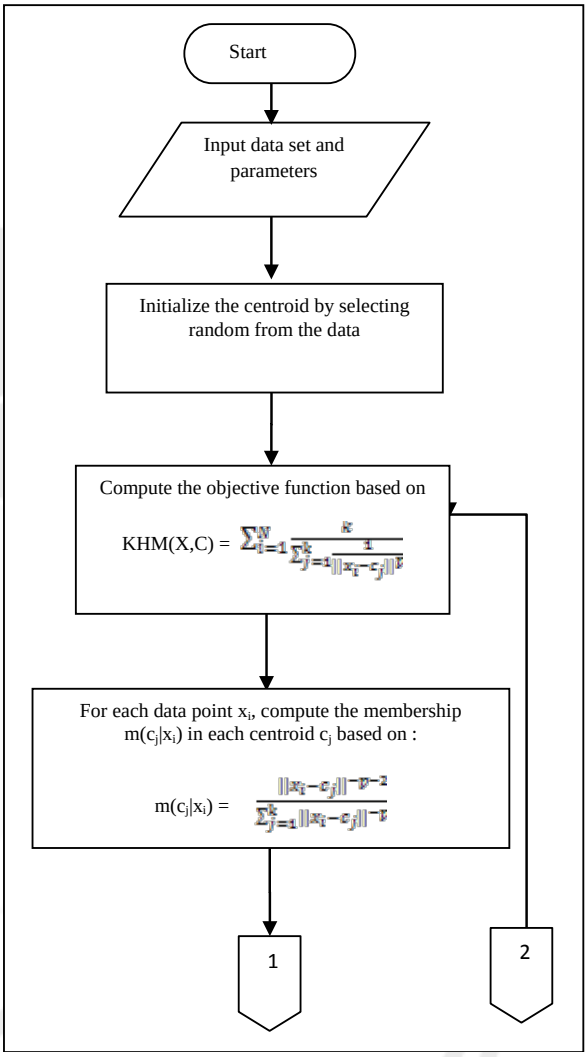


Figure 2. Flowchart KHM module sect.2

3. PARTICLE SWARM OPTIMIZATION

PSO method was introduced by Kennedy and Eberhart in 1995. PSO uses set of particles, where each particle represents a candidate solution, to explore solutions that allow for optimization problems. Each particle is initialized at random or heuristic, then the particles are allowed to "fly". At each step of optimization, each particle is allowed to evaluate the ability and the ability of particles in the vicinity. Each particle can store solution that produces the best value as one of the the best candidates solution for all the neighboring particles. PSO initialized with a random matrix production. The rows in the matrix is called a particle[2].

These lines contain variable values. Each particle will move according to distance and speed. Update particle velocity (velocity) with the equation 1 and its position based on the best solutions to local and global equation 2 [4].

$$V_i^{t+1} = \omega V_i^t + C_1 * R_1 * (P_i^t - X_i^t) + C_2 * R_2 * (P_g^t - X_i^t) \quad (1)$$

$$X_i^{t+1} = X_i^t + V_i^{t+1} \quad (2)$$

Variable i is the i-th particle in the flock, t is the number of iterations, V_i is the particle i velocity and X_i is the variable vector particles (eg position vector) of the i-th particle in N-dimensional problem.

P_i is the local best solution of the i-particles are obtained, and P_g is the global best solution of all particles in which P_i and P_g obtained based on the best fitness value [4]. R_1 and R_2 is a random number between 0 and 1, ω is the weight of particles called innertia weight, C_1 and C_2 are two constant numbers, often referred to as cognitive confidence coefficient [4].

PSO particle in the case of clustering is the matrix of data from each cluster centroid. PSO particle representation can be seen in Figure 3.

x_{11}	x_{12}	…	x_{1d}	…	x_{k1}	x_{k2}	…	x_{kd}

Figure 3. Particle representation

Where *k* is the number of clusters formed and *d* is the dimension of data. Fitness function used in this case is the objective function at the KHM algorithm [2]. PSO algorithm can be seen in Figure 4 – Figure 6.

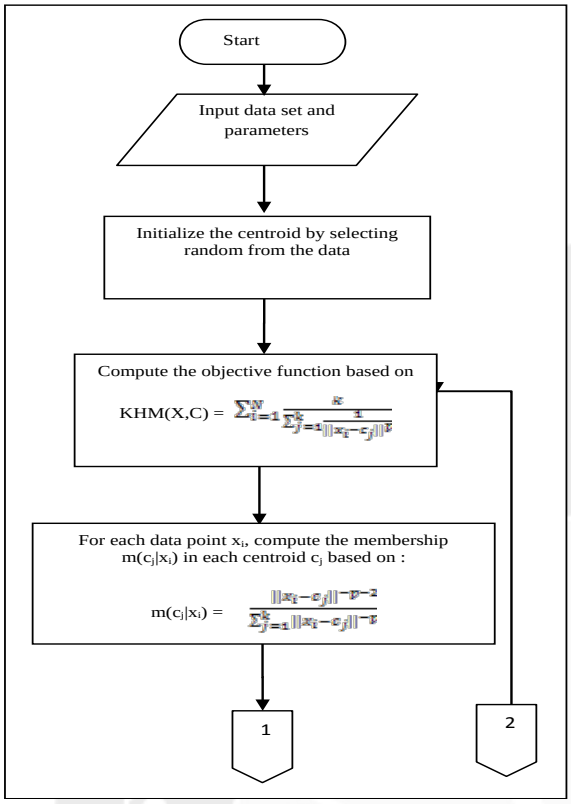


Figure 4. Flowchart PSO module sect.1

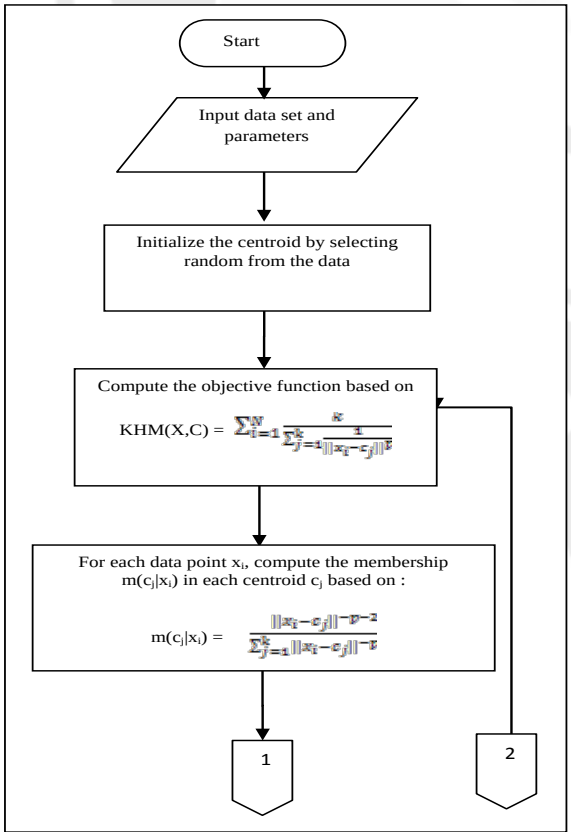


Figure 5. Flowchart PSO module sect.1

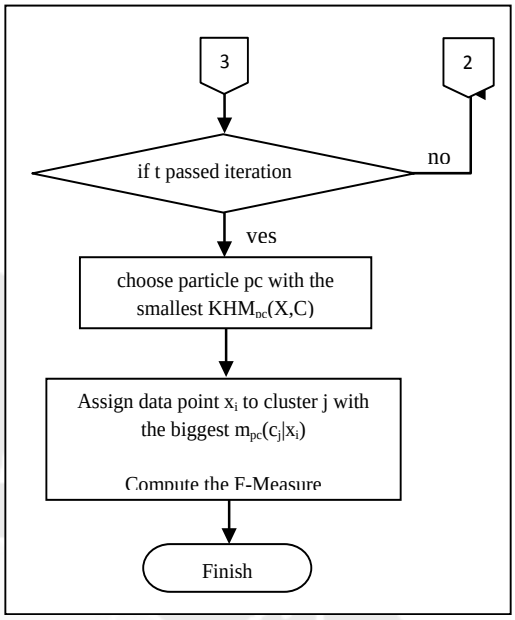


Figure 6. Flowchart PSO module sect.3

4. PSOKHM ALGORITHM

KHM is a clustering algorithm that can solve the problem of cluster centers initialization of K-Means algorithm. Although KHM can solve the initial problem, KHM still can not overcome local optima problem. To obtain the optimal solution, there is a stochastic algorithm called PSO, but the PSO algorithm has a problem at the speed of convergence. To overcome these problems, Fengqin Yang, Tieli Sun, and Changhai Zhang (2009) integrates the PSO algorithm with KHM, thus obtained hybrid clustering algorithm called PSOKHM. KHM PSOKHM algorithm using the four-times iteration of every eight generations of particles so that fitness values of each particle increases. Particle is a vector of real numbers of dimension k * d, where k is the number of clusters and d is the dimension of clustered data. Fitness function used for the algorithm is the objective function PSOKHM of KHM algorithm [2]. PSOKHM algorithm can be seen in Figure 7 and 8.

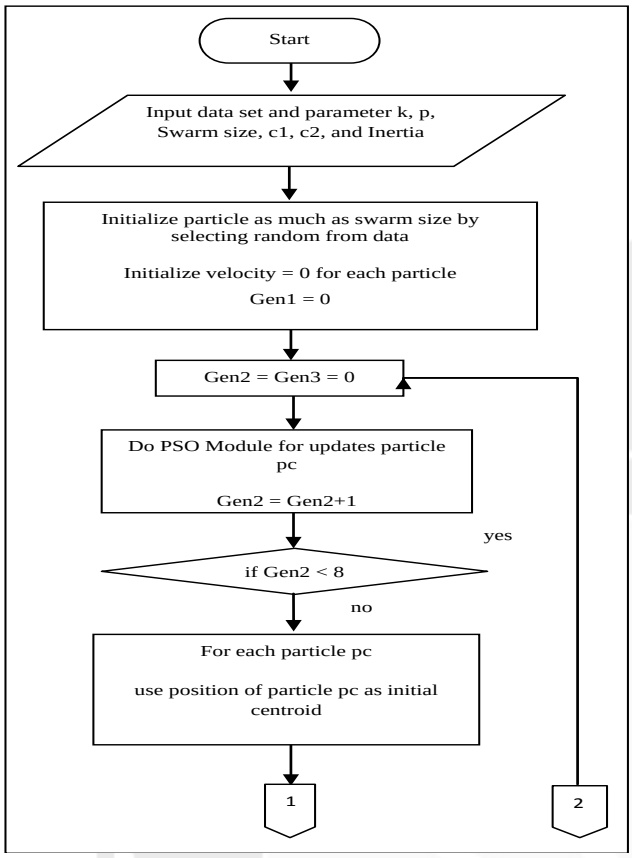


Figure 7. Flowchart PSOKHM module sect.1

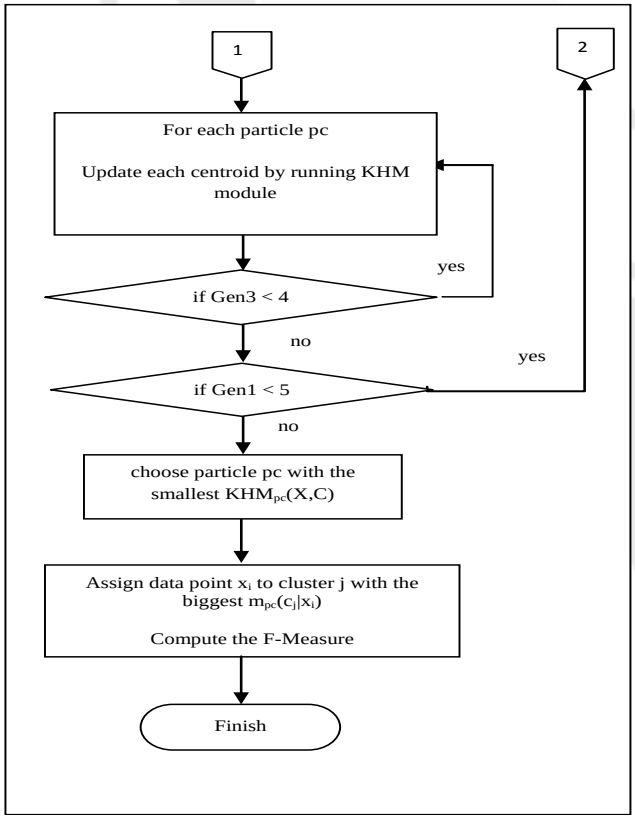


Figure 8. Flowchart PSOKHM module sect.2

5. EXPERIMENTAL RESULTS

Tests conducted on KHM algorithm, PSO, and PSOKHM with the following scenario:

1. Trial with parameter p = 2.5
2. Trial with parameter p = 3
3. Trial with parameter p = 3.5

Five data sets used as input for testing the system. The data sets used are Iris, Glass, Cancer, CMC, and Wine, where five data sets are stored in a similar file name with data extension. Data sets can be obtained from the website: ftp://ftp.ics.uci.edu./pub/machine-learning-databases/. Characteristics of each data set can be seen in Tablel 1. In addition to the five data sets that have been mentioned above, there are several input parameters that can be seen in Table 2.

P parameter value used to test the system is 2.5, 3, and 3.5. The value of k parameters depend on how many classes of each data set that can be seen in the column number of classes (k) in Table 1. While the other parameters value of iteration, Swarm size, c1, c2, and Inertia according to Table 2. Parameter values were chosen based on studies of PSO parameters selection by Shi and Eberhart [6].

Each algorithm is run 10 times for each data set, then the quality of clustering results from the three algorithms compared based on

1. Objective function value of KHM(X, C) is a sum of harmonic average between the data points with the centroid. The smaller the value of KHM (X, C), the better the quality of these clusters.
2. F-Measure is the value obtained from the measurement precision and recall of a class cluster results with the actual classes present in the input data. Precision and recall can be obtained by the following formula [2]:

$$\text{Precision } (i,j) = n_{ij}/n_j \quad (3)$$

$$\text{Recall } (i,j) = n_{ij}/n_i \quad (4)$$

Then the formula for calculating the value of F-Measure class *i* in cluster *j* is as follows [2]:

$$F(i,j) = \frac{(b^2+1).(p(i,j).r(i,j))}{b^2.p(i,j)+r(i,j)} \quad (5)$$

n_i is the number of data in class *i* are expected as a result of the query, n_j is the number of data in cluster *j* generated by the query, and n_{ij} is the number of elements from class *i* which in cluster *j*. To obtain a balanced weighting between precision and recall, used the value b = 1.

To get F-Measure value of the data set with the number of data n, the formula used is as follows:

$$F = \sum_i \frac{ni}{n} max_j\{F(i,j)\} \quad (6)$$

The greater the F-Measure, the better the quality of these clusters [2].

Algorithm is implemented using Matlab 7.0 on Intel Pentium Dual Core 1.86 GHz with 1 GB of RAM.

Table 1. Characteristics of data sets

Name of data set	No. of classes (k)	No. of Features (d)	Size of data set (n)
Iris	3	4	150
Glass	6	9	214
Cancer	2	9	683
CMC	3	9	1473
Wine	3	13	178

Using by 3 experiment scenarios, such as parameters p = 2.5, 3, and 3.5 so resulted, F-Measure and the running time of three algorithms which can be seen in Table 3 until Table 5. Value of the bold is the best value and in italics are the second best.

Table 2. Parameters value

Parameter	Value
p	2,5 , 3, dan 3,5
k	Depends on data set
Iteration	10
Swarm size	20
c1	1.49618
c2	1.49618
Inertia	0.7298

If carried out t test and ANOVA of the results of objective function KHM (X, C), F-Measure, and the running time of three modules that are contained in Table 3 – Table 5, obtained the following results:

a. Based on the confidence interval value of t test results on objective function KHM (X, C) in Figure 9, it was found that the results of the PSOKHM algorithm is better than KHM and PSO algorithm. From Figure 9 can also be seen that the difference results from the third objective function algorithms are not significant. This can be seen from the large P value.
b. Based on the confidence interval value of t test results on F-Measure in Figure 10, it was found that the results of the PSOKHM algorithm is better than KHM and PSO algorithm. Because F Measure is better if the value is larger, then the value of confidence intervals seen opposite. From Figure 10 can also be seen that the difference in F-Measure results of the three algorithms is not significant. This can be seen from the large P value.
c. Based on the confidence interval value of t test results on running time in Figure 11, it was found that the results of the PSOKHM algorithm is better than PSO algorithm, but even worse when compared with the KHM algorithm.

Table 3. Results from the KHM, PSO, and PSOKHM module on five data sets with p = 2.5

	KHM	PSO	PSOKHM
Iris			
KHM(X,C)	*148.904*	178.793	**148.876**
F-Measure	0.849	*0.859*	**0.871**
Running Time	**0.114**	7.895	*4.481*
Glass			
KHM(X,C)	*1193.53*	1467.35	**1182.752**
F-Measure	*1*	0	*0.537*
Running Time	0.534	**0.558**	*19.639*
	0.989	33.514	
Cancer			
KHM(X,C)	*58404.3*	70215.0	**58256.86**
F-Measure	*97*	25	**4**
Running Time	**0.851**	0.799	*0.850*
	0.356	40.459	*23.092*
CMC			
KHM(X,C)	*96201.4*	115027.	**96188.51**
F-Measure	*77*	103	**0**
Running Time	0.463	**0.482**	*0.477*
	2.678	110.740	*69.706*
Wine			
KHM(X,C)	*7533858*	793464	**75338461**
F-Measure	*5.310*	05.488	**.109**
Running Time	0.690	**0.705**	*0.694*
	0.939	22.185	*12.637*

Table 4. Results from the KHM, PSO, and PSOKHM module on five data sets with p = 3

	KHM	PSO	PSOKHM
Iris			
KHM(X,C)	*126.078*	155.417	**125.955**
F-Measure	0.868	**0.895**	*0.871*
Running Time	**0.092**	3.615	*2.050*
Glass			
KHM(X,C)	*1397.11*	2086.810	**1396.194**
F-Measure	*3*	0.549	*0.579*
Running Time	**0.579**	13.750	*8.004*
	0.346		
Cancer			
KHM(X,C)	*116341.*	147307.0	**115452.66**
F-Measure	*723*	55	**3**
Running Time	*0.800*	0.767	**0.826**
	0.326	17.582	*10.389*
CMC			
KHM(X,C)	*187018.*	240311.9	**186946.87**
F-Measure	*21*	8	0.464
Running Time	**0.481**	*0.469*	*25.814*
	1.270	45.456	
Wine			
KHM(X,C)	**104909**	1178075	*10491273*
F-Measure	**0406.35**	510.12	*12.42*
Running Time	**0.649**	0.639	*0.648*
	0.532	7.628	*4.342*

Table 5. Results from the KHM, PSO, and PSOKHM module on five data sets with p = 3.5

	KHM	PSO	PSOKHM
Iris			
KHM(X,C)	*109.823*	166.177	**109.694**
F-Measure	0.868	*0.873*	**0.874**
Running Time	**3.856**	7.707	*4.367*
Glass			
KHM(X,C)	*1881.275*	3511.627	**1856.299**
F-Measure	**0.580**	0.539	*0.575*
Running Time	**0.598**	35.464	*21.132*
Cancer			
KHM(X,C)	*237918.712*	313623.40	**234207.637**
F-Measure	0.827	9	0.826
Running Time	**0.771**	**0.870**	23.232
		40.382	
CMC			
KHM(X,C)	*380733.235*	551419.21	**380462.021**
F-Measure	**0.459**	4	0.449
Running Time	**7.831**	*0.459*	*69.746*
		113.454	
Wine			
KHM(X,C)	*154462516*	16466326	**14203924927.**
F-Measure	*26.98*	808.29	**48**
Running Time	*0.649*	0.635	**0.652**
	6.207	21.521	*12.451*

t-test results and anova for objective function
analysis of variance

'Source'	'SS'	'df'	'MS'	'F'	'Prob>F'
'Columns'	[1.9140e+017]	[2]	[9.5702e+016]	[0.0061]	[0.9939]
'Error'	[6.6016e+020]	[42]	[1.5718e+019]	[]	[]
'Total'	[6.6036e+020]	[44]	[]	[]	[]

t-test for KHM and PSO
h=1, Confidence Interval = -95873776.709, -57917170.179
t=test for KHM and PSOKHM:
h=1, Confidence Interval= 65180599.861, 100458741.116
t=test for KHM and PSOKHM:
h=1, Confidence Interval=141435716.368, 177994571.498

Figure 9. The results of t and Anova test to the objective function KHM (X, C)

```
*******************************************************
hasil uji t dan anova f-measure
*******************************************************
hasil uji anova :
tables =

    'Source'     'SS'             'df'    'MS'             'F'         'Prob>F'
    'Columns'    [3.0093e-004]    [ 2]    [1.5047e-004]    [0.0059]    [0.9941]
    'Error'      [     1.0719]    [42]    [     0.0255]    []          []
    'Total'      [     1.0722]    [44]    []               []          []


hasil uji t khm dengan pso :
h = 1 , confidence interval = 0.002534 0.003999
hasil uji t khm dengan psokhm :
h = 1 , confidence interval = -0.003802 -0.002331
hasil uji t pso dengan psokhm :
h = 1 , confidence interval = -0.007078 -0.005589
```

Figure 10. The results of t and Anova test to the F-Measure value

From the results of ANOVA test on running time there are significant differences between the KHM algorithm with PSO algorithm and PSOKHM, this can be seen by the P value approaches the value 0.

t-test results and anova for running time
analysis of variance

```
tables =

    'Source'     'SS'             'df'    'MS'             'F'         'Prob>F'
    'Columns'    [8.2100e+003]    [ 2]    [4.1050e+003]    [7.5462]    [0.0016]
    'Error'      [2.2847e+004]    [42]    [   543.9781]    []          []
    'Total'      [3.1057e+004]    [44]    []               []          []
```

t-test for KHM and PSO:
h=1, Confidence Interval = -33.075073, -32.851194
t=test for KHM and PSOKHM:
h=1, Confidence Interval=-19.015332, -18.874934
t=test for KHM and PSOKHM:
h=1, Confidence Interval=13.886324, 14.149676

Figure 11. The results of t and Anova test to running time

6. CONCLUSION

After a series of tests and analysis of the system's created, it can be concluded as follows:

a. PSOKHM algorithm can solve the problems of data clustering with better performance than the PSO and KHM algorithm if seen on the results of objective function KHM (X, C) and F-Measure.
b. PSOKHM not only increase the speed of convergence of PSO algorithm, but also helps KHM algorithm to move away from local optima.
c. When compared with KHM algorithm, PSOKHM algorithms require a longer time in the process of computing, so PSOKHM should not be applied if the time allowed very limited.

7. REFERENCES

[1] Cui, X., & Potok, T. E. (2005). Document clustering using Particle Swarm Optimization. In: IEEE swarm intelligence symposium. Pasadena, California.

[2] Fengqin Yang, Tieli Sun, Changhai Zhang. 2009, *An efficient hybrid data clustering method based on K-harmonic means and Particle Swarm Optimization*, Expert Systems With Applications 36, (pp. 9847-9853)

[3] Hammerly, G., & Elkan, C. (2002). Alternatives to the k-means algorithm that find better clusterings. In: Proceedings of the 11th international conference on information and knowledge management (pp. 600–607).

[4] Kennedy, J., & Eberhart, R. C. (1995). *Particle swarm optimization*. In Proceedings of the 1995 IEEE international conference on neural networks (pp. 1942–1948). New Jersey: IEEE Press.

[5] Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schutze. 2009. *Introduction to Information Retrieval*. Cambridge University Press.

[6] Shi, Y. H., Eberhart, R. C., 1998. Parameter Selection in Particle Swarm Optimization, The 7th Annual Conference on Evolutionary Programming, San Diego, CA.

[7] Ünler, A., & Güngör, Z. (2008). Applying K-harmonic means clustering to the partmachine classification problem. Expert Systems with Applications

[8] Zhang, B., Hsu, M., & Dayal, U. (1999). K-harmonic means – a data clustering algorithm. Technical Report HPL-1999-124. Hewlett-Packard Laboratories.

Implementation of Starfruit Maturity Classification Algorithm

R.Amirulah
Computer Vision, Video and Image Processing (CvviP) Research Lab
Faculty of Electrical Engineering
Universiti Teknologi Malaysia
81310 UTM Skudai, Johor
mr_rahman84@yahoo.com

M.M.Mokji
Department of Microelectronics and Computer Engineering (MiCE)
Faculty of Electrical Engineering
Universiti Teknologi Malaysia
81310 UTM Skudai, Johor
musa@fke.utm.my

Z.Ibrahim
Department of Mechatronics and Robotics (MER)
Faculty of Electrical Engineering
Universiti Teknologi Malaysia
81310 UTM Skudai, Johor
zuwairiefke@gmail.com

ABSTRACT

In this paper, an implementation of starfruit maturity classification algorithm based on YCbCr color space in an embedded system is presented. The maturity of the starfruit is classified based on new color feature, which is the hue between red and green color components denoted as m. The RG hue is derived and represented based on YCbCr because the input data to main processing is YCbCr format. Field Programmable Gates Array (FPGA) is selected to be the main processor of the system because of the capability of parallel execution on the algorithm and also because of its low power consumption. The image data is transferred into the FPGA through ITU-R 656 Decoder with YCbCr 4:2:2 color systems. Firstly, the system will segment the starfruit image in order to remove the background by using fixed threshold value. Then, index maturity classification is performed using simple rule-based classifier. In this work, classification accuracy of 90% was achieved and the result display on LCD.

Keywords

Real-time system; FPGA; Color features; CMOS; ITU-R 656 Decoder; YCbCr; LCD.

1. INTRODUCTION

Malaysia has been the largest exporter of starfruit in the world since 1989 [1].The biggest starfruit farm has also been setup in Selangor in 2002 [2]. It becomes a special production because the fruit is not only popular among Malaysians but also to the other peoples around the world. In 2008, 3648.9 metric ton which is about RM 25.5 million of starfruit was exported to various countries over the world such as Europe country (Netherlands, France, Germany, Canada), country from Middle East (Saudi Arabia, Iran, Bahrain, Turkey), and Asian country (Singapore, Hong Kong, Indonesia). Until Jun 2009, the export record shows that about 901.509 metric ton (RM14.5 million) totals starfruit was exported from Malaysia [3].

The qualities of the fruits exported from Malaysia are well known because the production of the fruit is controlled under Malaysia's Best food safety and quality regulation. All fresh producers must comply with the requirement set by the following standard where it was registered under Malaysian Standard fresh fruit (MS1127:2002 2nd Revision)[4].

In 2008, Malaysia's Best have reviewed the standard by including regulation on grading, packaging and labeling of agricultural product [5]. This regulation is aims for enhancing the quality of agriculture product.

Grade of the fresh starfruits can be divided into 3 which are Grade Premium, Grade 1 and Grade 2. From table 1, one of the important specifications for quality inspection is the maturity of the fruit. Quality inspection is a vital process to ensure only good qualities are being exported. Traditionally, evaluation of starfruit is performed by humans. These manual operations are time consuming; moreover the accuracy of this operation cannot be guaranteed. The starfruit quality inspections are based on its taste and physical appearance. Malaysia is acknowledged to have the best taste of starfruit amongst the importer countries compared to other exporter countries [6].

Starfruit maturity can be determined using the color of its skin. Based on FAMA rules for maturity grading, there are 7 different maturities called index 1 to index 7. Previously, the number of indices was 6 and being reviewed to 7 in June 2006 [7]. The six starfruit indices are shown in table 2. In this work, 6 indices version is applied as most of current practice is still based on the 6 indices.

Table 1. Grade specification for starfruit

Grade	Specification	Diffuseness
Premium	It must be the same verities, clean and fresh. The size and maturity is mostly equal. Without defect.	Maturities < 3% Freshness < 5% Defect < 3% Flaw < 3% Size Equality < 5%
1	It must be the same verities, clean and fresh. The size and maturity is mostly equal. Minimum defect.	Maturities < 5% Freshness < 5% Defect < 5% Flaw < 5% Size Equality < 10%
2	It must be the same verities, clean and fresh. The size and maturity is mostly equal. Without defect. Minimum defect.	Maturities < 10% Freshness < 10% Defect < 10% Flaw < 10% Size Equality < 10%

Table 2. 6 Indices of starfruit

Index	Explanation
 Index 1	**Dark green**. Fruit is immature. Not Suitable for harvesting.
 Index 2	**Green with a little yellow**. Matured. Suitable for harvesting for far export from sea.
 Index 3	**Green more than yellow**. Matured. Suitable for harvesting for far export from sea.
 Index 4	**Yellow green**. Far exporting using air transport.
 Index 5	**Yellow with a little green**. Still can be sent for far export using air transport.
 Index 6	**Yellow**. Not suitable for far exported. Local marketing.

There are few of similar systems have been designed for fruits such as apples [8], orange [9], papaya [10] and other. Since each fruits have its own characteristics, it is difficult to design a general machine for all fruits. So, a specific machine has to design to solve this problem especially for starfruit because starfruit has a complicated shape which is unique as it has five ridges forming a star shape while the other fruits only have a flat surface or rounded. Due to this unique shape, it is necessary to see the starfruit in single view to classify the maturity because the color changes on the starfruit surface are average for all surface ridges.

This paper propose an implementation of the 2 colors hue algorithm represented based on YCbCr color space in real-time system for quality inspection of starfruit using image analysis and vision technology. In the 2 section of this paper describe the overall system that will be design. The theory on the algorithm used in this work is in section 3. The system design is explained in section 4 including the experimental setup for the current system and the result of the experiment is discussed in section 5. The conclusion of this paper is on section 6.

2. OVERALL SYSTEM

In previous work, starfruit automation on quality inspection which is a computer based prototype machine vision system was designed based on color of the starfruit [11]. The algorithm for the system also used Hue parameter as the classifier where color recognition was established using multivariate discriminant analysis.

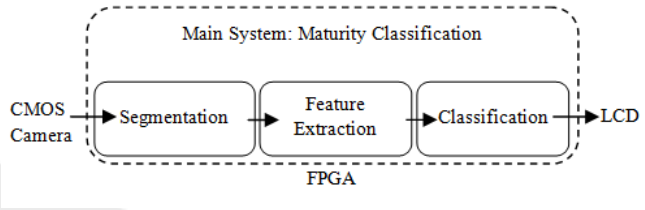


Figure 1. Overall system.

In this work, a simplified version Hue is considered, which is based on 2-dimensional color mapping (RG Color) [12]. The Hue of RG color is derived by using YCbCr color space component. YCbCr color space components are considered in this design because the data sent from camera into FPGA is in YCbCr data format based on video data transmission [13]. The color classification are differentiated into six maturity indices which is the rule produced by FAMA. This design can be divided into 3 sub-system which are input (Image Acquisition), main processing system (FPGA), and output (LCD). The overall system depicted in figure 1.

2.1 Image Acquisition

In this paper, CMOS sensor has been selected because of it characteristics which are high noise immunity and low static power consumption. The size of the active image grabbed is 640x480 pixels.

The image acquisition tool used in this work is a digital color camera (Olympus). Image captured by the camera is then transferred to FPGA via analogue video (AV) through the ITU-656 decoder.

2.2 Field Programmable Gates Array (FPGA)

A common configuration for very-high-volume embedded systems is the system on a chip (SoC) which contains a complete system consisting of multiple processors, multipliers, caches and interfaces on a single chip. SoCs can be implemented as an application-specific integrated circuit (ASIC) or using a field-programmable gate array (FPGA).

Implementation of the algorithm to system varies from Digital Signal Processing (DSP), Field Programmable Gate Array (FPGA) and Application Specific Integrated Circuits (ASICs) [14]. Considered the platform is the highest performances like ASIC but this type of system is hard to design and also too expensive. This is because ASIC design is not reprogrammable.

Table 3. Comparison between DSP, FPGA and ASIC

	DSP	FPGA	ASIC
Examples	TMS320, SHARC	Altera, Xilinx	-
Ease of Development	High	Medium	Low
Complexity & Cost	Low	Medium	Low
Power Consumption	Low	Low	Low
Performance	High	Medium	High
Reprogrammable	Yes	Yes	No

Most of the design is using DSP and FPGA system platform for the image processing. The implementation on DSP platform is outstanding compare to the FPGA for processing a single pixel of image. However, this design is most suits for FPGA platform due to the benefit of parallel execution and algorithm. The FPGA based hardware implementation profits especially from the high parallelism in the algorithm and the moderate number precision required to preserve the qualitative effects of the mathematical models. Furthermore, different variants can be supported on the same hardware by uploading a new programming onto the FPGA.

3. THE ALGORITHM

For this classification system design, Hue is used as the input features to the classifier. In previous work, simplified Hue (m_p) is used which is based on 2 color components [15]. The 2 colors Hue are represented by subtraction of red and green components as shown in Equation 1, where p are the number of pixels within the region of interest.

$$m_p = R_p - G_p \tag{1}$$

This paper representing the 2 colors Hue algorithm transform into YCbCr color space components. The YCbCr color space conversion formula to RGB color space is represented in Equation 2.

$$\begin{bmatrix} R_p \\ G_p \\ B_p \end{bmatrix} = \begin{bmatrix} 1.164 & 1.596 & 0 & -222.912 \\ 1.164 & -0.813 & -0.392 & 135.616 \\ 1.164 & 0 & 2.017 & -276.8 \end{bmatrix} \begin{bmatrix} Y_p \\ Cr_p \\ Cb_p \\ 1 \end{bmatrix} \tag{2}$$

From Equation 2,

$$R_p = 1.164Y_p + 1.596Cr_p - 222.912 \tag{3}$$

$$G_p = 1.164Y_p - 0.813Cr_p - 0.392Cb_p + 135.616 \tag{4}$$

By substituting Equation 3 and Equation 4 into Equation 1,

$$m_p = 2.409Cr_p + 0.392Cb_p - 358.528 \tag{5}$$

The m_p value in Equation 5 is the Hue of 2 colors, which is same with the subtraction of red and green colors in previous algorithm.

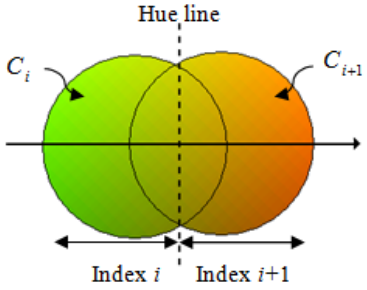


Figure 2. Classifier model.

The starfruit maturity classification process is based on the hypothesis where for each of the maturity index, certain area of the starfruit surface is supposed to have a distinctive range of the hue values that will differentiate between the six maturity indices. This hypothesis is described by the classification model shown in Figure 2. Based on this hypothesis, two parameters are involved, which are the starfruit surface area (Λ) measures in area percentage and hue (H) as a hue value that separates two adjacent maturity classes.

Figure 2 illustrates the hypothesis where two starfruits with adjacent maturity index are separated by a hue value. In this figure, the two starfruits are presented by their range of possible hue values and they are denoted by circle C_i and C_{i+1}. Figure 2 also shows that the two adjacent starfruits classes have partially overlap hue. This situation indicates that only a certain area on a starfruit can be used to determine its maturity index.

Because of the two adjacent starfruit classes in Figure 2 are only partially overlapped, total areas of the C_i and C_{i+1} on the left side of the hue line are always distinctive. Similar is the case on the right side of the hue line. As an example, if 70% of the C_i area lies on the left side of the hue line, total area of the C_{i+1} on the same side should be less than 70%. Thus, percentage area as well as the hue value that separates the two maturity classes were experimented in order to find the values.

As there are six maturity indices, there are five hue values that will separate the maturity indices. Thus, the five best values for the percentage area (desired areas) and the five best values for the hue (desired hues) were searched. The desired percentage areas are denoted as Λ_{di} and the hue values will be denoted as H_{di} where $i = 1, 2, 3, 4, 5$. Next, few rules was constructed to classify the starfruits into its maturity index.

Basically, the search for the Λ_{di} and H_{di} values are based on minimizing the class error denoted as Equation 6. The class error $E_{\Lambda,H,i}$ is denoted as equation 7. The minimum class error is used as it is a common measure of a quality estimator such as minimum square error (MSE) and minimum mean square error (MMSE) [16]. However, computation of the error in Equation 1 has been customized for the proposed classification model.

$$[\Lambda_{di}, H_{di}] = \arg\left[\min_{\Lambda=(1,2,3\ldots100)}\left(\arg\left[\min_{H=(-1,-0.01,-0.02\ldots1)}\left(E_{\Lambda,H,i}\right)\right]\right)\right] \tag{6}$$

$$E_{\Lambda,H,i} = \sum P\{\Lambda_{i,H} \le \Lambda\} + \sum Q\{\Lambda_{i+1,H} > \Lambda\} \tag{7}$$

In Equation 7, arguments P{.}=1 and Q{.}=1 if the arguments are true and set to zero if otherwise. Thus, $E_{\Lambda,H,i}$ is actually the total number of samples that do not satisfy conditions in equation 7 for certain values of area (Λ) and hue (H). Both samples in class C_i and C_{i+1} are included in the class error computation as the classification model in Figure 2 involves both classes. Specifically, $\Lambda_{i,H}$ is quantified based on Equations 8 as below. In words, $\Lambda_{i,H}$ will calculate the percentage area of a starfruit surface of class i that has hue value less than or equals to H.

$$\Lambda_{i,H} = \frac{number\ of\ pixels\ in\ C_i \le H}{total\ pixel\ of\ C_i} \times 100 \tag{8}$$

Based on the five chosen values in each Λ_{di} and H_{di} from the training process, the starfruit maturity classification can be achieved using simple rule-based classifier as shown in Figure 3. $\Lambda_{H_{di}}$ in the classification rules are computed based on Equation 9. The classifications start with comparing $\Lambda_{H_{d1}}$ with Λ_{d1} and stop when one of the rule arguments is true. If no argument is true, the starfruit will be classified as index 6.

$$\Lambda_{H_{di}} = \frac{number\ of\ pixel\ in\ C_i \le H_{di}}{total\ pixel\ of\ C_i} \times 100 \qquad (9)$$

4. DESIGN AND IMPLEMENTATION

4.1 FPGA Based Embedded System

Figure 3 show the general block diagram for the starfruit maturity classification design. Basically the system will be designed using hardware/software co-design technique [17]. The codes are then compiled to Verilog Hardware Description Language (HDL) [18]. A profiling will be done based on processing time for each stage. Verilog modules are design for all system architecture to benefit the parallel execution.

From the digital camera, the image data transfer to FPGA trough ITU-R 656 decoder [19]. The YCbCr 4:2:2 data store first on the SDRAM before the pixels data calls by VGA for display. In this case, SDRAM is used as buffer with first in first out (FIFO) type of data control. The data reads from SDRAM which is YCbCr 4:2:2 converted into YCbCr 4:4:4 before converted into RGB color space. The conversion of color space is done pixel by pixel for every frame.

After the color space conversion, the outputs are 10-bits RGB for the VGA display. In this stage, the 8-bits MSB of RGB output form color conversion take as an input for the main processing.

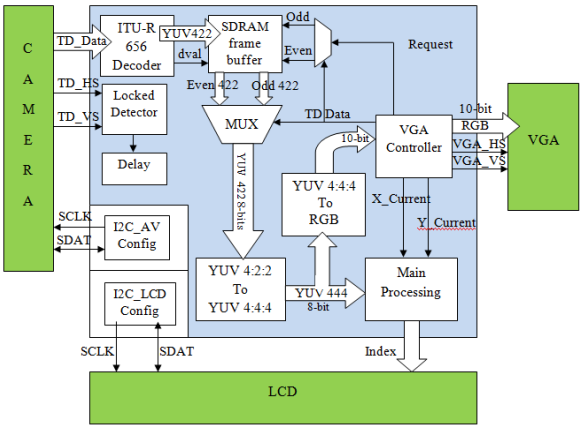


Figure 3. System architecture

4.2 Implementation of the Algorithm

The implementation of the algorithm is applied on main processing architecture by using Verilog HDL code for the FPGA. The implementation involved three stages of data flow, which are segmentation, feature extraction, and classification as shown in Figure 4. Also, note that this main processing is process a pixel of data in a time.

Once the sensors detect a starfriut, segmentation for the region of interest (ROI) is preceding and this segmentation is based on pixels value with fixed threshold value [20], which is if the value of Cb less than 115, the pixels is within the ROI. If the pixels values satisfy the segmentation rule, the system proceeds with m_p calculation otherwise the system proceed to the next pixel. m_p is calculated based on equation 5 with p is the pixels number within the ROI.

After the m_p calculation, the system will directly compare the m_p with the five fixed Hue value. If m_p value is less than the fixed hue values, the pixel is counts as H_m. Then, the $\Lambda_{H_{dm}}$ calculation will be processed where $m = 1,2,3,4,5$. After the $\Lambda_{H_{dm}}$ calculation, the system checks if one frame is done, otherwise the system will proceeds into next pixel data.

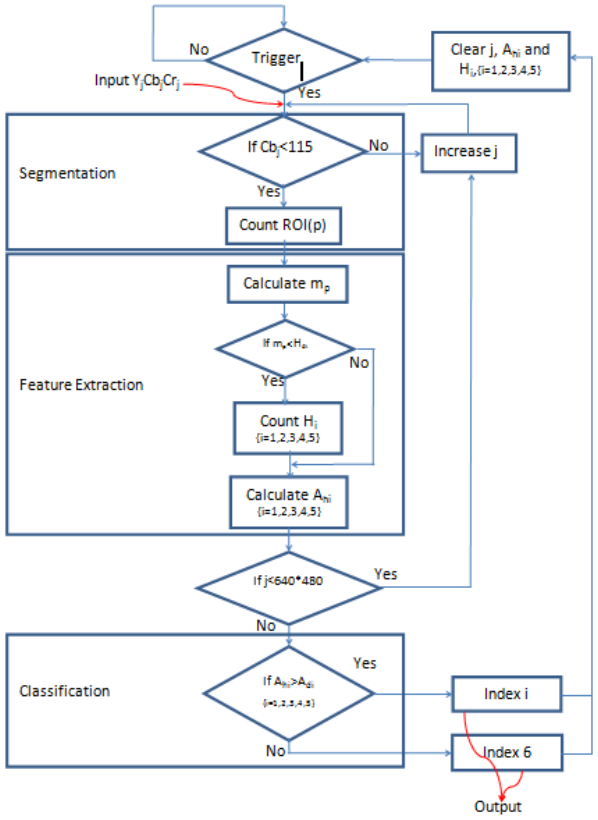


Figure 4. Main processing flow

Once a frame of image was processed, the value of $\Lambda_{H_{dm}}$ and Λ_{dm} are compared and the classification will stop if the argument is true. If there are no argument satisfied the comparison, the system will classified the starfruit as index 6. Finally, the output is sent to LCD for display.

5. RESULTS AND DISCUSSION

The algorithm has been successful implemented on the FPGA using Verilog HDL code. After translating the MATLAB code to Verilog code using Quartus II 8.0, the program loaded into FPGA board using USB-BLASTER that provided in DE2-70 Educational Board and the program immediately running on the FPGA board. This current system can be assumed as on-line system because the image loaded from the camera is a streaming video image 15 frame/second. The camera always send the image data frame into FPGA board and the image data always update even the image source from camera is a static image. This work was tested using starfruit image samples which are about 600 images with 6 different indices.

The classification results are shown in table 4. Based on the table, the results show that the error are highiest at index 4 and index 5. At index 4, most of the image was classified as index 3. This is because the starfruits image samples for index 3 and index 4 were almost same in color. The error in index 5 classified as index 6 because the image samples for index 5 were more orange color for the starfruit images.

6. CONCLUSION

Vision technology is rapidly used in fruits inspection especially for it skin or surface color. With a new system designed for the inspection purpose, the quality of the fruit can be classified its maturity and separated easily for the export purpose to the around the world. This embedded system will advance the user by lowering cost of design and improving the performance in productivity. Compare to the computer based system; this system is less costly, expected performance and also low power consumption.

7. ACKNOWLEDGMENT

This work funded by Ministry of Science and Technology Innovation Malaysia (MOSTI) under vot 78367 done at Universiti Teknologi Malaysia.

Table 4. Results

Index	% Correct	% Error
1	100	0
2	89	11
3	97	3
4	81	19
5	74	26
6	99	1
Average	90	10

8. REFERENCES

[1] Maria J. Dass and Gabrielle Chaik, “Star Attraction”, Cover Stories in Sun2Surf, 23 Sep. 2005.

[2] Rosliwaty Ramly, “Selangor Starfruit Valley”, Bernama, 28 March 2005.

[3] Mohd Hamirol Ab Hamid, FAMA Johor Malasia, Report “Quantity and Export Value of Fruits by Country for 2004 until June 2009”, hamirol@fama.gov.my, 14 Oct 2009.

[4] Malaysian Standard, “Specification for Fresh Carambola MS1127:2002”, as for 31 Disember 2009. http://www.standardsmalaysia.gov.my/.

[5] News, “Fama didik pengguna kenali 3P”, Sinar Harian (Pahang), 31 Oct 2009.

[6] Jonathan H. Crane, “Commercialization of Carambola, Atemoya, and Other Tropical Fruits in South Florida”, Proceedings of the Second National Symp.on Exploration, Research & Commercialization, page: 448-460, 1993.

[7] Manual Quality Series – “CARAMBOLA”, www.famaxchange.org, June 2006.

[8] Q. Yang, “ Automatic detection of patch-like defects on apples” , Fifth International Conference on Image Processing and its Applications, 1995, Page(s):529-533, 4-6 Jul 1995.

[9] M. Recce, J . Taylor, A.Piebe, G. Tropiano, “ High speed vision-based quality grading of oranges”International Workshop on Neural Networks for Identification, Control, Robotics, and Signal/ Image Processing, 1996, Page(s):136-144, 21-23 Aug. 1996.

[10] S. Limsiroratana, Y. Ikeda, “ On image analysis for harvesting tropical fruits, Proceedings of the 41st SICE Annual Conference, Volume 2, Page (s) : 1336-1341 v ol.2, 5 - 7 Aug. 2002.

[11] M.Z. Abdullah, A.S. Fathinul-Syahir, B.M.N. Mohd-Azemi, “Automated inspection system for colour and shape grading of starfruit (Averrhoacarambola L.) using machine vision sensor”, Transactions of the Institute of Measurement and Control 27, 2 (2005) pp. 65-87.

[12] M.M. Mokji, S.A.R. Abu Bakar, “Starfruit Grading Based on 2-Dimensional Color Map”, Regional Postgraduate Conference on Engineering and Science (RPCES 2006), Johore, 26-27 July.

[13] Keith Jack, Video Demystified: A Handbook for the Digital Engineer, Fifth Edition, Elsevier Inc. 2007

[14] Jefri Mustapa, Ahmad Zuri Sha’ameri, Muhammad Mun’im Zabidi, “Reconfigurable Embedded Vessel Classification System for High Frequency Telemetry Application”, Faculty of Electrical Engineering, UniversitiTeknologi Malaysia, Skudai 81300, Johor, Malaysia, 2009

[15] Musa Bin Mohd Mokji, “Features Contruction for Starfruit Quality Inspection”, PhD Thesis, UTM, January 2009.

[16] Chang, C.I.; Du, Y.; Wang, J.; Guo, S.M. and Thouin, P.D. “Survey and Comparative Analysis of Entropy and Relative Entropy Thresholding Techniques”. IEEE Proceedings on Vision, Image and Signal Processing. December 2006. 153(6): 837-850.

[17] Edwards, M.D., Forrest, J., Whelan, A.E, “Acceleration of software algorithms using hardware/software co-design techniques”, Journal of Systems Architecture 42, 697-707, 1997.

[18] Mohamed Khalil Hani, “Starter’s Guide to Digital Systems VHDL & Verilog Design”, Pearson Prentice Hall, 2007.

[19] Terasic, “DE2-70 User Manual, TV Decoder Specification”, www.terasic.com, Pages 52-54, 2009.

[20] Rafael C. Gonzalez and Richard E. Woods, “Digital Image Processing”, 2nd. Ed. New Jersey: Prentice Hall. 2002.

Improving Choquet Integral Agent Network Performance by Using Competitive Learning Algorithms

Handri Santoso
Nagaoka University of Technology
Kamitomiokamachi 1603-1, Nagoka
Niigata 940-2188 JAPAN
Telp: +81-258-47-9375
handri.santoso@mis.nagaokaut.ac.jp

Shusaku Nomura
Nagaoka University of Technology
Kamitomiokamachi 1603-1, Nagoka
Niigata 9402188 JAPAN
Telp: +81-258-47-9375
nomura@kjs.nagaokaut.ac.jp

Kazuo Nakamura
Nagaoka University of Technology
Kamitomiokamachi 1603-1, Nagoka
Niigata 9402188 JAPAN
Telp: +81-258-47-9375
nakamura@kjs.nagaokaut.ac.jp

ABSTRACT

Choquet Integral Agent Networks (CHIAN) could realize the operational means embedding existing partial, qualitative knowledge and gray box representations for the information fusion mechanisms. However, the effectiveness of CHIAN for solving real world problems has some limitation, first, if it has to handle large number of input channels (n) where CHIAN requires (2^n) subset values to determine fuzzy measure (w). Second, if human knowledge should be embedded to each input channel of CHIAN agent. From the practical standpoint, these conditions are highly infeasible and are hardly implemented in many realistic problems. Thus, this study proposed a method to decompose the input patterns of network structure into some categories. The results showed the effectiveness of the proposed method.

Keywords

Choquet Integral Agent Network, Competitive Leaning Algorithms. Macroscopic and Microscopic Information, Fuzzy Measure.

1. INTRODUCTION

It is natural that human should seek to design and build machines that can recognize pattern. From automated speech recognition, fingerprint identification, optical character recognition, DNA sequence identification, and much more, it is clear that reliable accurate pattern recognition by machine would be immensely useful. For some problems, such as speech and visual recognition, our design efforts may in fact be influenced by knowledge of how these are solved in nature, both in algorithms we employ and in the design of special-purpose hardware [1]. Pattern recognition is the scientific discipline whose goal is the classification of objects into a number of categories or classes. The degree of difficulty of the classification problem depends on the variability in the feature values for objects in the same category relative to the difference between feature values for objects in different categories. The variability of feature values for objects in the same category may be due to complexity, and may be due to noise. There is no single classifier that works best on all given problems as explained by No-free-lunch theorem. Determining a suitable classifier for a given problem is however still more an art than a science.

Many problems in real world are always accompanied by imprecision or uncertainty factors. Sometimes the information is far from complete, but a precise decision is required. In such situation, human can make a correct decision, while computer requires complex calculation to make a mathematical model of the problem. In this way, the development of flexible and intelligent information processing mechanisms based on human thinking mimicry is required for solving real-life problems. In this context, new types of various computing methods were proposed to cope with such aspects of information processing for the real world problems. These methods are called “soft computing” as a whole, and they are respectively attributed with specific features and play supplementary roles to each other. The developments of fuzzy logic, probabilistic reasoning, neural network and evolutionary algorithm motivated the researchers to explore the possibilities of building more humanlike machines using these tools. The synergisms of these tools might also improve the performance of the overall system to a great extent. It is a partnership in which each of the partners contributes a distinct methodology for addressing problems in its domain. In this perspective, the principal contributions of fuzzy logic, probabilistic reasoning, neural network and evolutionary algorithm are complementary rather than competitive [2].

As intelligent information processing mechanisms various soft computing methods have been introduced in recent years. Fuzzy Sets, Fuzzy Measure and Integrals, Neural Networks and Genetic Algorithms were introduced for coping with complexity, unexpected change, and incomplete knowing of the real world. The correlations between these issues were introduced by [3] as shown in **Fig.1**. This figure shows the triangle of these three issues to be approached facing to the real world problems.

In this figure the symbols, i.e., *C*, *U* and *I*, stand for abbreviations of “complexity”, “unexpected change”, and “incomplete knowing” respectively. To cope with these issues, the descriptions follow as;

1) for complexity of the real world problems limited human ability might employ macroscopic description using conceptualization, approximation and summarization,
2) for unexpected change of time-spatial states of the real worlds adaptable human ability might employ flexible processing using elastic, plastic and / or floating thinking,
3) for incomplete knowing about the real world experienced human ability might employ subjective knowledge like belief, intuition and emotion.

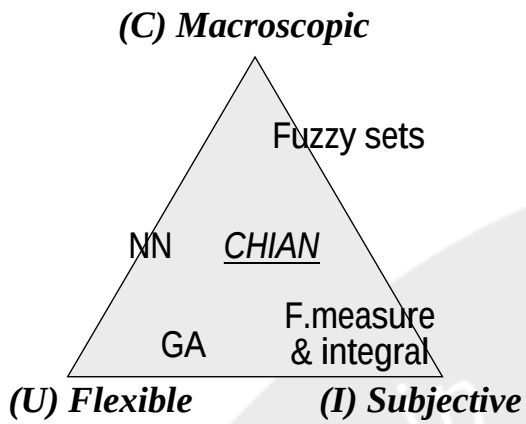


Figure 1. Features of flexible human intelligence and soft computing methods

Allocating the existing representative methods into this framework, they can be distinguished by their own features which can work supplementary to each other. For advancing the computing methodology these methods have to be compromised at conceptual levels.

In [4] a fusion of neural network concepts and Fuzzy Measure and Integral concepts as a flexible information fusion mechanism is studied. The developed method is based on the fact that theory of neural networks is designated to represent hidden mechanisms of information transformation in human, biological or natural phenomena. While it is attributed to high flexibility, however, it could not make them to comprehend the real mechanism in most cases. On the other hand, theory of fuzzy sets or theory of fuzzy measure and integrals are useful for enhancing their macroscopic comprehension of complex systems in the real world. For the problems being appropriate to neural networks approaches, he proposed to introduce a Choquet integral mechanism in the framework of fuzzy measure at every neuron unit instead of simple weighted sum mechanisms. Then the units may work more flexibly and more meaningfully as intelligent human information processors. As the results the proposed information fusion mechanisms, i.e., Choquet Integral Agent Networks (CHIAN) could realize the operational means embedding existing partial and qualitative knowledge and gray box representations for the information fusion mechanisms. The allocation of CHIAN is embedded in **Fig.1**.

CHIAN has beneficial features to realize the operational means embedding existing partial and qualitative human knowledge, flexible model of the network, and gray box representation. However, the effectiveness of CHIAN for solving real world problems has some limitation, first, if it has to handle large number of input channels (n) where CHIAN requires $\left(2^n\right)$ subset values to determine fuzzy measure (w). Second, if human knowledge should be embedded to each input channel of CHIAN agent. From the practical standpoint, these conditions are highly infeasible and are hardly implemented in many realistic problems. In this case, the input patterns of network structure have to be decomposed into some categories. Though, the input grouping is not easy task, in the most cases, some prior knowledge and/or trial several experiments are required in this process. In this way, considering several issues for improving CHIAN as classifier will be described as follow

- Some parts of the network are able to learn from the input data by exploring the input pattern.
- The network is able to memorize the characteristic of training data, by exploit the locality information of input pattern.
- Automatic generations of the hidden agents of CHIAN structure are developed by learning the input characteristics.

2. GENERATION HIDDEN UNITS OF CHIAN BY USING COMPETITIVE LEARNING ALGORITHMS

Learning from data is one of the essential capabilities of machine learning for copying human ability when interpreting the patterns of data. Clustering is one of the most primitive mental activities of humans, used to handle the huge amount of information they receive every day. Processing every piece of information as a single entity would be impossible. Thus, human tend to categorize entities into clusters. Each cluster is then characterized by the common attributes of the entities it contain. Clustering is a method which allows discovering similarities and differences among patterns and deriving useful conclusions about the patterns of data. In view of this fact, the clustering method is employed for categorizing inputs of CHIAN.

Among the clustering methods, competitive learning algorithms are one of the effective unsupervised learning which can learn data without a teacher. This method has capability to learn the input pattern without known or assumed number of clusters. The learning adjustment is confined to the single cluster center that is most similar to the pattern currently being presented. As the result, the characterizations of previously discovered clusters that are unrelated to the current pattern are not disrupted.

Competitive learning algorithms employ a set of representatives w_j, $j = 1,..., J$. The goal is to move each of them to regions of the vector space that are "dense" in vectors of X. The representatives compete with each other when a new vector $x \in X$ is presented to the algorithm. The winner of this competition is the representative that lies closer to x. Then the winner is updated so as to move toward x, while the losers either remain unchanged or are updated toward x but at a much slower rate. The competitive learning scheme may be stated as follows;

- (A) Initialization $w_1 \leftarrow$ randomly $x \in X$ as initial cluster centers.
- REPEAT
  - Present randomly input vector $x \in X$
  - (B) Determine the winning representative w_c
  - (C) IF $\|x\text{-}w_c\|$>threshold AND $J<J_{max}$

    $J=J+1$

    $w_J=x$

- ELSE
  (D) parameters updating

$$w_j(t)=\begin{cases} w_j(t-1)+\eta\left(x-w_j(t-1)\right) & \text{if } w_j \text{ is the winner} \\ w_j(t-1) & \text{otherwise} \end{cases}$$

- UNTIL convergence has occurred
- (E) Eliminated cluster which have only a few members
- Updating the rest of clusters

In this algorithm, the initial cluster is given randomly by using a value of vector *X*. Then, the input vectors are presented in a different order, i.e., $x_1, x_2, \cdots, x_N$, $x_5, x_8, \cdots, x_{N-6}$ for each iteration. The determination of the winning representative (part (B)) is carried out using the following rule;

$$\left\|x-w_c\right\|=\min_j\left\|x-c_j\right\| \quad (1)$$

$\|\ \|$ is similarity or dissimilarity measure, such as Euclidean distance, Mahalanobis distance, or etc. In case of similarity measure used, the min operator in this preceding relation is replaced by the max operator. The numbers of created representative (part (C)) are limited by $J_{\max}$ and given threshold depending on the application at hand. The updating of the representatives (part (D)) is carried out by equation:

$$w_j(t)=\begin{cases} w_j(t-1)+\eta\left(x-w_j(t-1)\right) & \text{if } w_j \text{ is the winner} \\ w_j(t-1) & \text{otherwise} \end{cases} \quad (2)$$

where η is the learning rate and takes values in [0,1]. According to this algorithm, the losers remain unchanged. On the other hand, the winner w_j moves toward x. The size of the movement depends on η. In the extreme case where η =0, no updating takes place. On the other hand, if η =1, the winning representative is places on x. For all other choices of η, the new value of the winner lies in the line segment formed by $w_j(t-1)$ and x. After the all cluster are found, then check whether the cluster have enough members or not. The clusters which don't have enough members will be eliminated. Finally, the unassigned vectors are presented to the algorithm and are assigned to the appropriate cluster.

For a better illustration, let us consider an artificial data example where the data consist of eight measurements as shown in **Table 1**. Therefore the categorizations data by competitive learning create two representatives. The data distribution and created representative data generating by competitive learning is shown in **Figure 2**.

Table 1 Example of artificial data

Data	1	2	3	4	5	6	7	8
Variable 1	2.0	6.0	5.0	2.0	1.0	5.0	3.0	2.0
Variable 2	5.0	4.0	3.0	2.0	4.0	2.0	3.0	3.0

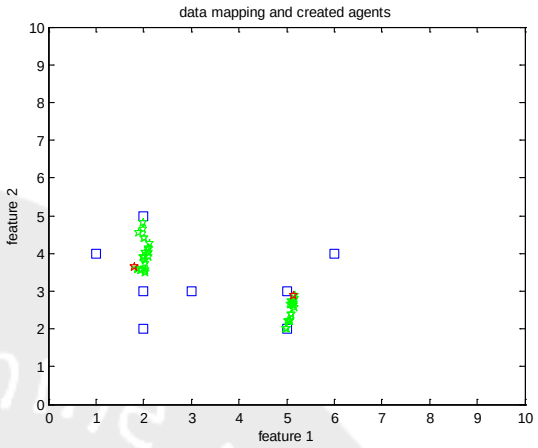


Figure 2. Using artificial data, the process for created two representatives is generated by competitive learning. The data is represented by blue square, while green and red pentagram represent the process of updating the representatives and the final value of the representative, respectively.

3. ARCHITECTURE OF CHIAN WITH COMPETITIVE LEARNING ALGORITHMS (CHIAN-CL)

To reduce subset of determined fuzzy measure, competitive learning algorithm is employed to categorize the input of CHIAN and the created representative's data are set as connection strength of hidden units in CHIAN structure. This hybrid method, i.e., unsupervised and supervised learning called CHIAN-CL is proposed in this study for improving CHIAN performance. Architecture of the proposed network is shown in **Fig. 3**. The structure of CHIAN-CL consists of three layers, i.e., input layer, competitive learning hidden layer, and output layer. Each layer consists of units, i.e., input units, hidden units and output units. The *i*-th input unit receives normalized external input $u_i \in [0,1]$ $(i=1,\cdots,N_I)$ to the network and outputs it as inputs to hidden units. Output $x_{Ii} \in [0,1]$ of the *i*-th input unit is given by:

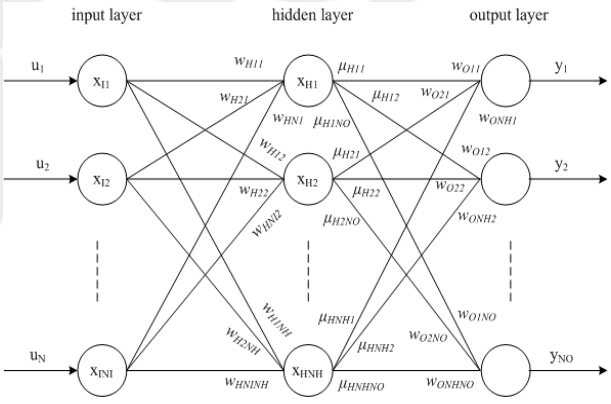


Fig. 3. Architecture of CHIAN with Competitive Learning Algorithms

$$x_{Ii} = u_i ; \left(i = 1, \cdots, N_I\right) \tag{3}$$

Output of hidden units is normalized by membership function $\left(\mu_{Hjk}, j = 1, \cdots N_H ; k = 1, \cdots, N_O\right)$. The Output of the *j*-th hidden unit is given by

$$\mu_{Hjk} = f\left(x_{Hj}, options\right) \in \left([0,1]\right) \tag{4}$$

$$x_{Hj} = \sqrt{\sum_{i=1}^{N_I} \left(w_{Hij} - x_{Ii}\right)^2} \quad j = 1, \cdots, N_H \tag{5}$$

where $w_{Hij} \in \Re$ represent the characteristics of input patterns which relates with connection strength of input units and the *j*-th hidden unit generating by competitive learning and *options* is the parameters of memberships function. A membership function is essentially a curve that defines how each point in the input space is mapped to a membership value (or degree of membership) between 0 and 1. The various types of membership function are normally used including triangular, trapezoidal bell shaped, Gaussian curves, polynomial curves and sigmoid function [61]. The output of hidden units become inputs of output units, these output unit provide the networks output

$$y_k = (c)\int \mu_{jk} \cdot dw_{jk} = \sum_t \left[\mu_{jk(t)} - \mu_{jk(t-1)}\right] \cdot w_{jk,t} \in \left([0,1]\right); k = 1, \cdots, N_O \tag{6}$$

where $w_{jk} \in [0,1]$ is a collection of fuzzy measure that respecting of input pattern of the *k*-th output unit. The given desired system output is denoted by d_k. In order to obtain a network that produces output y_k with respect to input u_i, the values of w_{Ojk} should be determined such that they will minimize the following error function:

$$E = \frac{1}{2} \sum_k \left(d_k - y_k\right)^2 \tag{7}$$

Then, the values of w_{Ojk} are modified iteratively by:

$$\Delta w_{Ojk} = -\eta \frac{\partial E}{\partial w_{Ojk}} \tag{8}$$

where $\eta \in (0,1)$ is positive constant. The calculation of $\partial E / \partial w_{Ojk}$ is performed by the back-propagation algorithm [5].

4. THE EXPERIMENTS AND RESULTS

4.1 Iris Problem

The Iris flower dataset is a popular multivariate dataset that was introduced by R.A. Fisher [6] as an example for discriminant analysis. The dataset consists of 50 samples from each of three species of Iris flowers, i.e., Iris setosa, Iris versicolor and Iris virginica. The three examples of Iris flower is shown in **Fig.4**. Four features were measured from each sample; they are the length and the width of sepal and petal. The iris data distribution is shown in **Fig. 5**. It shows that the Iris setosa is linearly separable from the other two; the latter are not linearly separable from each other.

In this study, classification of dataset is performed by the proposed method, i.e. CHIAN with competitive learning algorithm. In the initial stage, the input data is categorized by competitive learning to generate units in the hidden layer. The position of the created units is shown in **Fig. 6**. After the units of hidden layer are created by competitive learning, the data is transformed by (5) and then the result is normalized by membership function for preparing as input of output layer. The results of the transformed data are shown in **Fig. 7**. The classification accuracy is achieved in 97%, 86%, and 80% for class one, two and three, respectively as shown in **Table 2**.

(a) (b) (c)

Figure 4. Type of iris flower
(a) I. setosa (b) I. versicolor (c) I. virginica

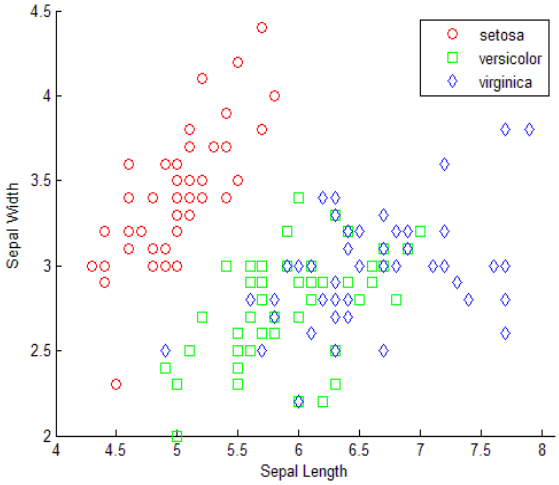


Figure 5. Distribution data of Iris

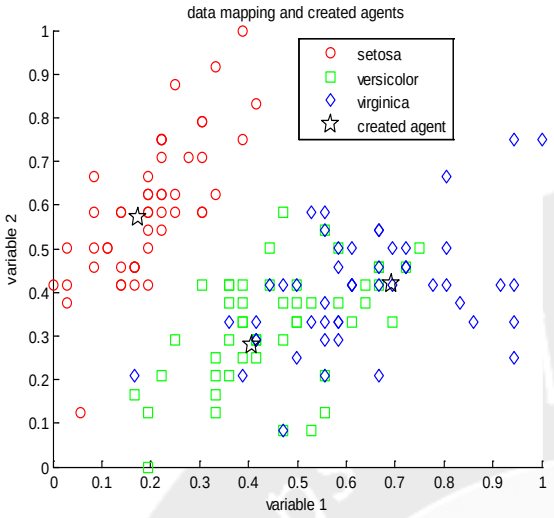


Figure 6. The created unit of hidden layer relates to the first and second variables.

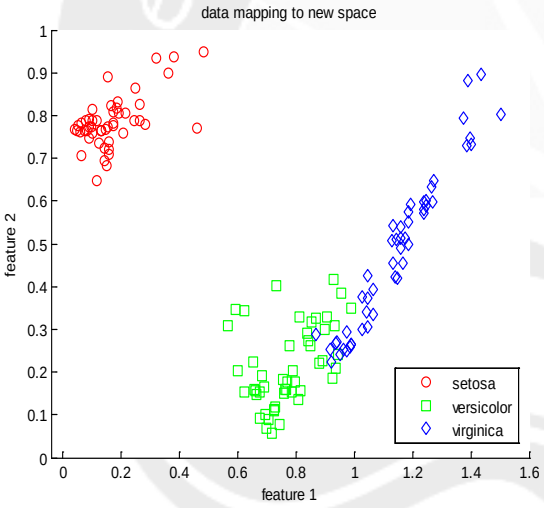


Figure 7. Distribution of data in hidden layer relates to the first and second units.

4.2 Wine Problem

Wine data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivated [6]. The dataset consists of 59, 71 and 48 samples of class one, two and three, respectively.

The analysis determined the quantities of 13 constituents found in each of the three types of wines, i.e., :

1) Alcohol
2) Malic acid
3) Ash
4) Alcalinity of ash
5) Magnesium
6) Total phenols
7) Flavanoids
8) Nonflavanoid phenols
9) Proanthocyanins
10) Color intensity
11) Hue
12) OD280/OD315 of diluted wines
13) Proline.

The same process is done for wine data problem. The created representatives wine data are shown in **Fig. 8**. The transformed data is shown in **Fig. 9**. The classification result of wine problem is 92%, 92% and 88% for class one, class two and class three, respectively, as shown in **Table 2**.

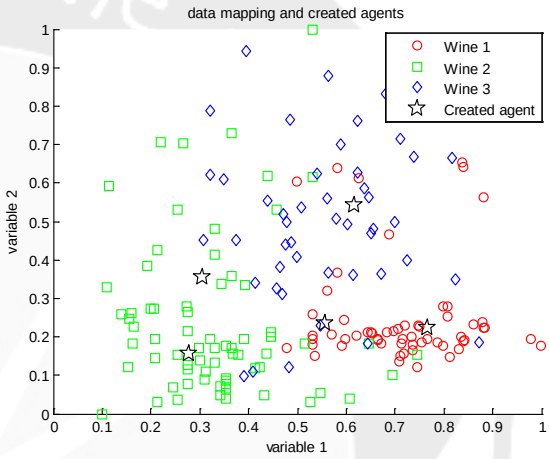


Figure 8. The created unit of hidden layer relates to the first and the second variables.

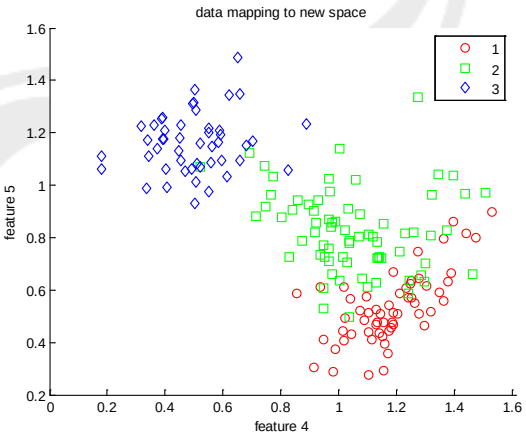


Figure 9. Distribution of data in hidden layer relates to the second and the third units.

Table 2. The classification results of CHIAN-CL

Dataset	Class 1	Class 2	Class 3
Iris	97%	86%	80%
Wine	92%	92%	88%

5. CONCLUSIONS

CHIAN has a drawback if it has to handle huge amount features of data. In addition, each agent (unit) of CHIAN should have a meaning function i.e., embedding human knowledge is required before constructing the CHIAN for solving the problem, instead of black box function. CHIAN-CL as an improvement of the CHIAN is proposed in this study. CHIAN-CL explore characteristic of input pattern that improves capability of CHIAN for solving real world problems. Consequently, embedding human knowledge to the CHIAN structure becomes easier. The created hidden units of CHIAN-CL which have capability to learn characteristics of the input pattern vector have dual benefit, i.e., reduce dimensionality and feature extraction capability.

The Iris and Wine dataset is one of the popular problems in which the data distribution is overlapping between classes. The experiment results show the effectiveness of the proposed method for solving these problems. Therefore, the proposed method is promising to solve the problem where the conventional CHIAN suffers from that problem conditions.

.

6. REFERENCES

[1] R. Duda, P. Hart, and D. Stork, "Pattern Classification," 2nd ed., New York, Wiley, 2000.

[2] The Berkeley Initiative in Soft Computing website, http://www-bisc.cs.berkeley.edu/BISCProgram/History.htm

[3] K.Nakamura, "From neural networks to Choquet integral agent networks," The 3rd Czech-Japan Seminar on Data Analysis and Decision Making under Uncertainty, 2000.

[4] Nakamura, K.,"A Scheme for information fusion by Choquet Integral Agent Networks," Eighth IFSA Congress, pp 954 – 958, 1999.

[5] Rumelhart, D.E., and Wiliam, R.J., "Learning Representations by Error Propagation," Nature 323, pp. 533-536, 1986.

[6] Asuncion, A & Newman, D.J. (2007). UCI Machine Learning Repository, Irvine, CA: University of California, Department of Information and Computer Science, http://www.ics.uci.edu/~mlearn/MLRepository.html.

Improving Food Resilience with Effective Cropping Pattern Planning Using Spatial Temporal-Based Updated Pranata Mangsa

Kristoko Dwi Hartomo
Faculty of Information Technology
Satya Wacana Christian University
Jl. Diponegoro 52-60, Salatiga
50711, Indonesia
kristoko@gmail.com

Sri Yulianto J.P.
Faculty of Information Technology
Satya Wacana Christian University
Jl. Diponegoro 52-60, Salatiga
50711, Indonesia
sriyulianto@gmail.com

Krismiyati
Faculty of Information Technology
Satya Wacana Christian University
Jl. Diponegoro 52-60, Salatiga
50711, Indonesia
blesschris@gmail.com

ABSTRACT
Pranata mangsa has been widely used in the society especially those living in rural area in Java. The global warming has caused climate changes affecting the use of pranata mangsa to be not valid anymore. This study will design a new model of pranata mangsa for determining an effective cropping pattern to support the national food resilience using spatial temporal base. The design of spatial temporal-based updated pranata mangsa system will be able to solve the problem caused by the climate changes. It could reduce the failure risk and increase the production and local food availability. Apart from that, it could also reduce the risk of nutrition and food vulnerability.

Keywords
pranata mangsa, spasial temporal, food vulnerability

1. INTRODUCTION
As a big and populated country, Indonesia is faced with a complex challenge in fulfilling the food needs. This phenomenon has triggered the food resilience policy as its central issue and the main focus of its agricultural development. Two main components of for realizing food resilience are improving the food needs along with the population increase and work opportunity for the society to obtain a decent income so for accessing the food. This food resilience policy could support the national food stability.

The main problem for realizing the food resilience is due to the fact that the growth of food demand is faster than its supply. The fast growing demand is a result of population, economic and buyer power growth and also the taste changes. On the other hand, the national food production capacity remains stagnant due to a competition in empowering the land resources. Besides, it is affected by the stagnancy of land productivity growth and the availability of agricultural wmorkers. The imbalance of demand growth and national production capacity tends to increase the national food supply obtained from import activity. This import dependency is closely related to national food supply stability [1].

At the moment, national food resilience is not strong yet. There are still problems in most of the food resilience aspects such as the policy which is not consistently applied, inappropriate food management, and the weak anticipation taken for disaster either in rainy or dry season. If there is no acceleration done with the condition and development pace happening right now, then it will not take longer time to suffer from a significant deficit of rice production [2].
Farmers often suffer from losses resulted from the crop failure due to draught or flood. As it happened in Kerawang Regency, almost 13,000 ha ricefield are damaged and failed due to draught. Meanwhile, there are 25,000 ha are failed to harvest due to flood [3].

A part from that, recently it is common to hear that water in a dam is shrinking causing need for electricity and irrigation could not be met. The core of the problem is the fact that this phenomenon often recoccurs as a result of ignoring or not learning from the past. One of the solutions for this problem is by having a good water management which is dependent to the climate condition.

Those natural phenomena are difficult to control and modify unless it is in a small scale. For optimizing those climate phenomena, information about climate condition especially its chance of having extreme climate such as long draought and flood and also climate forecast should be known as early as possible. This is for avoiding or minimizing the impacts resulted from those extreme climate [4].

Pranata mangsa is a traditional method for Javanese in forecasting the weather based on natural phenomenon. Threfore, the user of this method should “remember” (in Javanese : titen), when to plant and when to crop. The accuracy level of this traditional forecast is often bias as there might be some missing natural indicators due to some natural destruction.

Modern weather and climate forecast has not given any optimum result yet. Therefore, it is important to develop any method covering the instruments, modelling technique and also improving the human resources. Another effort for improving the accuracy level is by integrating the traditional method which is local and modern method which is already global. Integrating those two systems is not as easy as it seems to be and it should be continuously and carefully thought [4].

The traditional method of Pranata Mangsa has been really proved its existence and there are still some parties using it especially those people in rural area. The global warming has caused the climate changes and pranata mangsa use becomes invalid [4].

2. OBJECTIVES

This study aims to design a new model of pranata mangsa for determining effective cropping pattern for supporting the national food resilience. Besides, It aims to reduce the crops failure risk, increase production and local food availability and to reduce the food and nutrition vulnerability Figure 1.

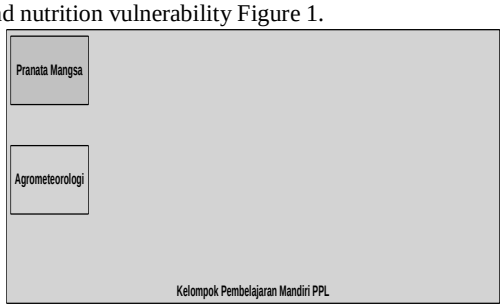


Figure 1. The position of updated pranata mangsa

3. REVIEW OF LITERATURE

3.1 Traditional Weather and Climate Forecast.

There are some traditional weather and climate forecasts such as Pranata Mangsa in Java, Kala in Sunda, Porhalaan in Batak, and Wariga in Bali. This study focuses more on pranata mangsa.

"Pranata mangsa" is from Javanese language. Pranata means procedure and mangsa means season. Mataram Kingdom, Sultan Agung created a Javanese Calendar by changing the calculation system of Saka year which is based on the moon revolution and its movement towards the earth just like the Hijriyah year. However, the year number follows the year number of Saka. He succeeded on integrating the method of Islamic and Javanese (Hindhu) [4].

Javanese calendar contains pranata mangsa. It is closely related to human characters, good day for trading, having business, wedding, moving house or when they should do a fasting day such as sanger, taliwangke, samparwangke, sarik agung, dhendhan kukudan, etc. Pranata mangsa is also used for stating to plant, harvest, and to plant crops.

Pranata mangsa in this study covers season division (mangsa), number of days, farmer activities, the seen characteristics (natural signs) in each of the seasons (Figure 2). The 365 days are divided into twelve seasons or known as "mangsa" in Javanese. Each season is different in its lenght ; Kasa (first): 41 days (23 June – 2 August), Karo (Second): 23 days (3 August -26 August), to Sadha (the twelfth): 41 days(14 May-22 June) (third circle) [4].

Figure 2. Pranata mangsa (Source: Ki Hudoyo Doyodipuro, modified : www.xentana.com/java/calendar.htm)

Those twelve seasons are classified into four general season (first circle) : they are dry season (88 days), labush (first transition : 95 days), rainy (94/95 days), and mareng (second transition: 88 days). Farmer activities for each season rotates anti clockwise (second circle). It starts from first season with planting the crops, second season for plant growth until the twelfth season of harvesting in the rice field. Apart from farmer activities, pranata mangsa also gives the characteristics or natural phenomenon for each season. An example could be seen in the first season (22 June – 2 August), the natural phenomenon is that the wind from the north east to the south west, high temperature, small fountain, falling leaves, grasshopper and insects laying thier egg.

Using pranata mangsa, farmers could plan when they have to start planting and when they are going to harvest. One example is that farmers could start planting paddy in the sixth and seventh seasons which are on November 10 – February 3. In those seasons there will be eind from the west to the east, damp temperature, cold, frequent flood and rain, rambutan and mangsoteen starts to reap especially in the sixt season. In this season birds are diffcult to get their food. Rice harvest could be predicted to take place in the tenth, eleven, and twelfth seasons. The naural phenomena in these seasons are strong and constant wind from the soouth east, little rain, birds starting to build their nest, and hot temperature [4].

3.2 Modern Weather and Climate Forecast

1980s is the starting point of modern weather and climate forecast development especially in Indonesia. This forecast is often represented in forecast model either deterministic or statistic. To make forecast model needs many data and complex analysis which often creates a particular problem. With the computing development, data analysis along with its complicated mathematical calculation is not a problem any longer [4].

For designing and developing climate and weather forecast needs surface data collection such as precipitation, temperature, dampness and pressure. It also needs data from far sensing such as NOAA (National Oceanic Atmosphere Administration) and GMS (Geostationary Meteorology Satellite). Sattelite data could record an area with a wide observation at simultaneously collected in one data scene. Therefore it could be used for observing climate and weather globally. Furthermore, it has high temporal resolution

which could be obtained every hour or days. Weather and climate forecast using sattelite data has been developd by LAPAN [4].

Many kinds of modelling technique have been used as well ranging from the simplest one to the most complicated one. Generally , weather forecast modelling uses deterministic approach while season and climate forecast model uses statistic approach. There are several stockastic (statistic) models developed in Indonesia such as time series modeal (ARIMA, winter-additive, transfer function), fourier regression, fractal analysis, trend surface analysis, neural network, transformasi wawelet, MARS, ItsMARS, dan analisis regresi (e.g. Dupe 1999; Haryanto 1999; Boer et al. 2000; Haryoko 1997; Zifwen 1999; Andriansyah 1998 dalam Sutikno, 2002). Meteorological and Geophysical agency uses probability method, harmonic series and analog method for forcasting the climate in Indonesia (Gunawan et al, 2001).

Many analyisis techniques and model development have been done for increasing the forecast accuracy. Table 1 shows some weather and climate forecast along wiith the climate changes which has been developed in Indonesia until 2006 [4].

Tabel 1. Weather and climate forecast and climate changes have been developed in Indonesia till 2006 [3]

No.	Nama Model	Instansi Peneliti
1	Model prediksi curah hujan berdasarkan anomali suhu permukaan laut (SPL)	LAPAN
2	Model pemantauan dan prediksi gerak ITCZ	LAPAN
3	Global Circulation Model (GCM)	LAPAN, IPB
4	Echam, Hirlam, RSM, MM5, CLARK dan ARPS	BPPT
5	Limited Area Model (LAM)	LAPAN
6	Model Prakiraan Musim	BMG
7	Model prediksi anomali suhu permukaan laut pasifik	ITB

There are some weather and climate changes forecasts adopted to develop in Indonesia. LAPAN has tried to develop a weather and climate forecast model using ITCZ model, anomali sea water surface temperature and General Circulation Models (GCM) CSIRO 9 level (Adiningsih et al, 2000). Nowadays, a model using GCM data which is global has been developed to forecast weather or climate which is local using downscaling technique [5].

3.3 Spatial Data Model

Spatial data is used and analysed using computer or known as spatial digital data take them as a model. Economic and Social Comminssion for Asia and the Pasific (1996) defines data model as a logic set or rules and characteristic of a spatial data. Data model represent the relation of real world and virtual world [6].

There are two models in spatial data, raster data model and vector data model. Both of them have different characteristics and the use depends on the dat input and the output. This model is a representative of the geographical objects recorded which could be reognised and processed by computer. Chang (2002) divides the vector data model into several parts (could be seen in Figure 3). It will be explained in the following section.

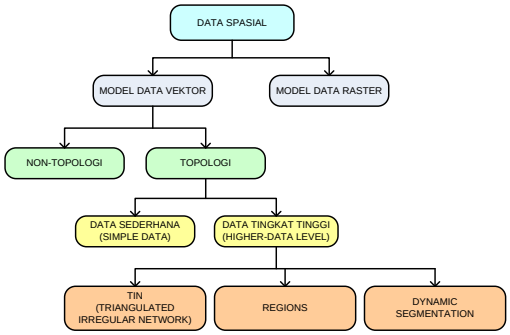


Figure 3. Spatial data model classification [6]

4. ANALYSIS RESULT AND DISCUSSION OF SYSTEM DESIGN

4.1 System Analysis

4.1.1 Functional Need

The system for making updated pranatamangsa model to determine the effective cropping planning with spatial model will be in form of agricultural areal map. It has detailed specification as follows :

- Could be used for inputting climatology data such as temperature, precipitation , and air humidity.
- Could be used for inputting agricultural production data such as kind of production plan, and production result)
- Could determine an object for finding its relation to other objects.
- Could determine an area to analyze based on a particular object .
- Could generate neighbourhood object relationship.
- Could show a map containing information of an area division in a research object, climate of a particular area, cropping pattern which is appropriate to the the climate in that particular area.

4.1.2 Process Analysis

The processes in this system are as follows :

1st Process : **Setting Up**, a process for determining data type used as a a parameter in the system and its relationship with the data in the map scale.

2nd Process : Generating **Neighbour**, a process for finding objects around an object inputted by the user. Neighbour Object are located in a range inputted by the user. The output of this process will be saved in the neighbour data.

3rd ProcessProses 3: Generating **Association Rule**, a process of object neighbour data relationship processing into association rule based on minimun support and confidence inputted by users. The output of this process will be saved in data rule.

4th Process: Reporting, a proses of showing implementation result containing information of areal division of the research object, condition of the climate in a particular area, and cropping pattern which is appropriate to the condition of that area.

4.2 System Design

4.2.1 Use case diagram

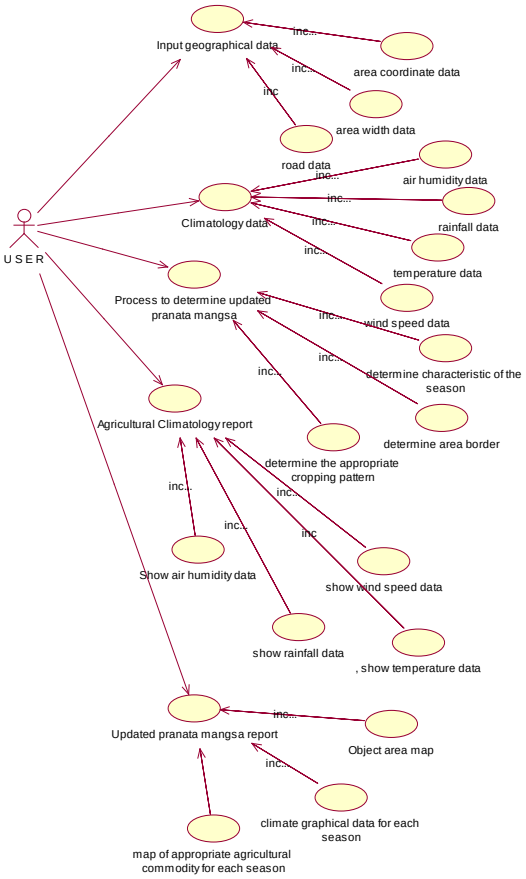


Figure 4. Use case diagram

4.2.2 Class diagram

Class describes states (attribute/property) of a system. It also offers a service for manipulationg of that states (method or function) Figure 5 shows the class diagram design for the updated pranata mangsa system for relating the data to produce a map of information about agricultural map, commodity of each area and its production, climate condition for each season and the most appropriate for a particular season.

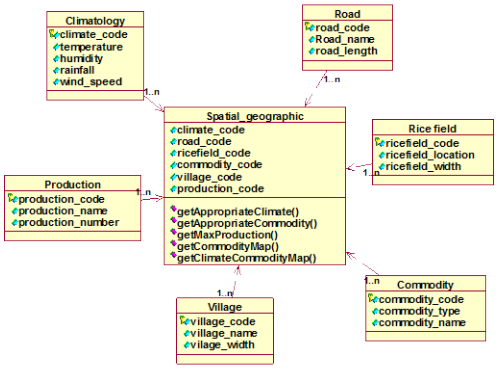


Figure 5. Class diagram

4.2.3 Spatial Data Base

Spatial data base is map of research object. This data base has several tables described as follows:

- Climatology Table
- Road Table
- Village table
- Rice field table
- Commodity table
- Production result table
- Geography table

4.2.4 System Architecture

The application is built using MapServer as CGI program and has the following architecture : browser (client) sends a request (through internet or intranet) to the web server in a spatial request (location [x,y], click cursor or layer status). The request is then sent to the application server and MapServer (CGI Program). After that, MapServer will read the MapFile, map data, and external data (if there is any and needed) to form a suitable picture as requested. Having the picture file renderred, this image file will be sent to the web server and then to forwarded to the browser client. The application architecture using Mapserver could be seen in Figure 5.

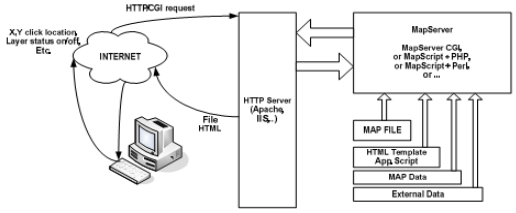


Figure 5. System architecture

4.2.5 Spatial Modelling using raser data model

Raster data representing thematic map could be derived from data analysis result. The activity carried out are as follows [6] :

- Classify the sattelite image to produce land cover .
- Classify the value of the multispectral data into a particular category (such as vegetation type) and assigning value to that category.
- Geoprocessing operation combined with many sources such as surface, raster , and vector data.
- Using agricultural climatology raster data as the input for producing cropping pattern appropriacy map using updated pranata mangsa as seen in Figure 6.

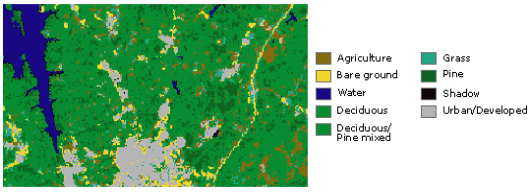


Figure 6. Raster model for planning effective cropping pattern.

4.2.6 System Interface

It is expected from the system built that it will solve the problem of climate change in pranata mangsa. In addition, it will be able to produce spatial based information which is more accurate and kind of commodity which is appropriate to the climate condition as seen in Figure 7.

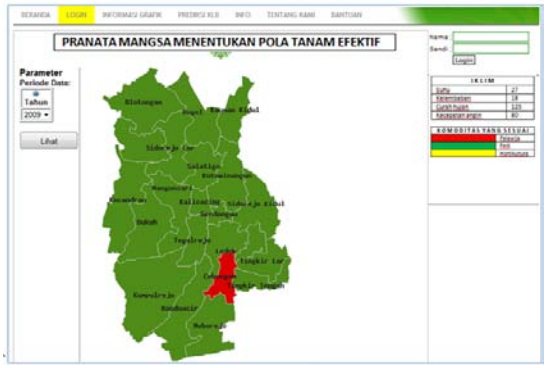


Figure 7. Interface of updated pranata mangsa system for planning effective cropping pattern

5. CONCLUSION AND SUGGESTION

The design of updated pranata mangsa system could solve the problem emerging as a result of climate changes in the current pranata mangsa.

The system is expected to be built and implemented to overcome the current problem so that it could produce spatial based information which is more accurate and the commodity which is suitable to the condition of the area.

6. REFERENCES

[1] Kepala Badan Litbang Pertanian Departemen Pertanian, Makalah Simposium Nasional Ketahanan dan Keamanan Pangan pada Era Otonomi dan Globalisasi, Faperta, IPB, Bogor, 22 November 2005.

[2] M. Hasan, Makalah Pengantar Falsafah Sains (PPS702) Program Pasca Sarjana / S3, Institut Pertanian Bogor, 28 November 2006.

[3] _____, Sekitar 30 Persen Sawah di Pantura Jabar Dilanda Banjir. Kompas, 22 Februari 2004.

[4] Sutikno, Makalah Pengantar ke Falsafah Sains (PPS702) Sekolah Pasca Sarjana / S3 Institut Pertanian Bogor, Mei 2004.

[5] B.S. Tedjakusuma, Adiningsih, Kajian pemanfaatan informasi cuaca dan iklim di Indonesia, Prosiding Lokakarya Sehari. LAPAN Jakarta, hlm 25-35, 2000.

[6] D. Gumelar, Document license: Copyright IlmuKomputer.Com, 2003-2007.

[7] D. Gunawan, Soetamto, Nuryadi, Heru, Prakiraan jangka panjang di badan meteorologi dan geofisika.. Di dalam M.A Ratag et al (Penyunting). Prediksi Cuaca dan Iklim Nasional. Prosiding Temu Ilmiah LAPAN, Bandung, hlm 51-59, 2001.

[8] Pranata mangsa: http://www.geocities.com/ sekar_jono/pramang.htm, 16 Mei 2004.

[9] The Javanese Calendar : www.xentana.com/ java/calendar.htm, 12 Mei 2004.

[10] Sutikno, Penggunaan Regresi Splines adaptif Berganda untuk Peramalan Indeks ENSO dan Hujan Bulanan, Tesis S2 IPB (Tidak dipublikasikan), 2002.

[11] Chang, Kang-Tsung. *Introdcution To Geographic Information Systems*. New York: McGraw-Hill, 2002.

[12] Economic and Social Comminssion for Asia and the Pasific. *Manual on GIS for Planner and Decision Makers*. New York: United Nations, 1996.

[13] M. Karimariyanti, D. Darmantoro, D.S. Kusumo, *Seminar Nasional Aplikasi Teknologi Informasi (SNATI) ISSN: 1907-5022 Yogyakarta, 16 Juni 2007,*

Knowledge Based System in Defining Human Gender Based on Syllable Pattern Recognition

Muhammad Fachrurrozi
Informatics Engineering, Computer Science, Sriwijaya University
Jalan Joko Atas No 23
Palembang, Indonesia
+62-85213355478
fachrur@yahoo.com

ABSTRACT
Registration process for event or activity participant becomes important moment because the given information will become reference to give attribute and making a decision to the participant. Gender is one of important attribute to be given. Native people in an area or a place have characteristics in giving name to their children. Usually, name can represent his gender: man or women. By Knowledge based system and word pattern recognition his name, we can get relative conclusion or suggestion about his own gender.

Keywords
Gender, knowledge based, syllable pattern recognition.

1. INTRODUCTION
By information technology advancing, almost event or activity implement the technology to get faster and more accurate result. The events or activities give a responsibility to the registrar to fill the registration form. In the registration process, the given information seldom occurs incorrectly. The incorrect information may occur from system or long time registration process. In many online registration (e.g. mail.yahoo.com) have already given another form in word pattern recognition based on given name to be used in defining unique ID suggestion.

Gender is one of human identity. This gender only have 2 (two) answer, man or woman. Thus, the incorrect information will be effect to next attribute which will be given to him. Every native people in an area or a place have name identity which are used to call or as a difference communication identity person to another person. In one area or place have characteristic how parent give their children name. For example in Palembang, person who has name “Yanti”, or “Tuti”, or “Santi” tend to a women gender. Thus, name with suffix ‘ti” tend to women gender with defined percentage.

Pattern recognition is already used by many researchers, in image or text form. In that way, they can get some conclusions to define another decision or to give some suggestions. In 2008, Xinyong do research about A Method for Evaluating the Sensitivity of Signal Features in Pattern Recognition Based on Neural Network. He used this algorithm to create a criterion function for evaluating the feature sensitivity [3]. And word pattern recognition also gives us some temporal conclusion, while he is a man or a woman by using neural network algorithm.

2. HUMAN GENDER IDENTITY
Gender is defined by FAO as ‘the relations between men and women, both perceptual and material. Gender is not determined biologically, as a result of sexual characteristics of either women or men, but is constructed socially. It is a central organizing principle of societies, and often governs the processes of production and reproduction, consumption and distribution’ [1]. Despite this definition, gender is often misunderstood as being the promotion of women only. However, as we see from the FAO definition, gender issues focus on women and on the relationship between men and women, their roles, access to and control over resources, division of labour, interests and needs. Gender relations affect household security, family well-being, planning, production and many other aspects of life [1].

Gender roles are the ‘social definition’ of women and men. They vary among different societies and cultures, classes, ages and during different periods in history. Gender-specific roles and responsibilities are often conditioned by household structure, access to resources, specific impacts of the global economy, and other locally relevant factors such as ecological conditions [1].

Gender relations are the ways in which a culture or society defines rights, responsibilities, and the identities of men and women in relation to one another [1].

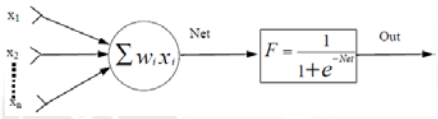
3. CHARACTERISTIC OF A NAME
A name is an important thing for people because of many reasons. A name represents as an identity and as a subject difference beside as an object of people to communicate each other. In giving the name for their children, parent have some reasons such as cultural reason, social status, religion, their hometown, or taken from famous people. The name also could be considered to a certain gender. i.e., the person who lives in Palembang may have the name such as “Santi”, “Fitriyanti”, “Tuti”, are tend to classify to woman gender. Otherwise, the name such as “Firman”, “Lukman”, “Lukman”, are classified to man gender.

The name may consist more than one word and every word may contain more than one syllable. Certain syllable also could be considered as identity, i.e. the name with suffix “ti”, “ni”, “na” are tend to woman gender. Otherwise the name with suffix “to”, “di”, “man” are tend to man gender.

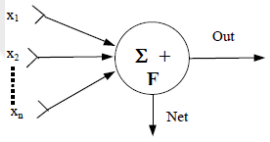
4. KNOWLEDGE BASED SYSTEM

Learning is an inherent characteristic of the human beings. By virtue of this, people, while executing similar tasks, acquire the ability to improve their performance. Machine learning can be broadly classified into three categories: i) Supervised learning, ii) Unsupervised learning and iii) Reinforcement learning. Supervised learning requires a trainer, who supplies the input-output training instances. The learning system adapts its parameters by some algorithms to generate the desired output patterns from a given input pattern. In absence of trainers, the desired output for a given input instance is not known, and consequently the learner has to adapt its parameters autonomously. Such type of learning is termed 'unsupervised learning'. The third type called the reinforcement learning bridges a gap between supervised and unsupervised categories. In reinforcement learning, the learner does not explicitly know the input-output instances, but it receives some form of feedback from its environment.

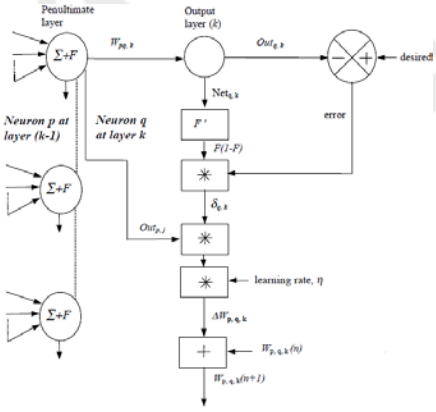
4.1 The Back-propagation Training Algorithm



(a) A typical neuron used in Back propagation algorithm



(b) Schematic representation of the neuron shown in (a)



(c) Training of a weight, Wp,q,k at the output (kth) layer.

Figure 1. Attributes of neurons and weight adjustments by the back propagation learning algorithm[3]

The back-propagation training requires a neural net of feed-forward topology. Since it is a supervised training algorithm, both the input and the target patterns are given. For a given input pattern, the output vector is estimated through a forward pass on the network. After the forward pass is over, the error vector at the output layer is estimated by taking the component-wise difference of the target pattern and the generated output vector. A function of errors of the output layered nodes is then propagated back through the network to each layer for adjustment of weights in that layer. The weight adaptation policy in back-propagation algorithm is derived following the principle of steepest descent approach of finding minima of a multi-valued function.

Typical neurons employed in back-propagation learning contain two modules (vide fig. 1(a)). The circle containing Σ wi xi denotes a weighted sum of the inputs xi for i= 1 to n. The rectangular box in fig. 1(a) represents the sigmoid type non-linearity. It may be added here that the sigmoid has been chosen here because of the continuity of the function over a wide range. The continuity of the nonlinear function is required in back-propagation, as we have to differentiate the function to realize the steepest descent criteria of learning. Fig. 1(b) is a symbolic representation of the neurons used in fig. 1(c).

In fig. 1(c), two layers of neurons have been shown. The left side layer is the penultimate (k –1)-th layer, whereas the single neuron in the next k-th layer represents one of the output layered neurons. We denote the top two neurons at the (k-1)-th and k-th layer by neuron p and q respectively. The connecting weight between them is denoted by wp,q,k. For computing Wp,q,k(n+1), from its value at iteration n, we use the formula presented in expression[2].

$$\delta = F'*(t\arg et - Out) = Out(1-Out)(t\arg et - Out) \quad \text{..... (4.1)}$$

$$\Delta w_{p,q,k} = \eta \delta_{q,k} Out_{p,j} \quad \text{..... (4.2)}$$

$$w_{p,q,k}(n+1) = w_{p,q,k}(n) + \Delta w_{p,q,k} \quad \text{..... (4.3)}$$

where:

$w_{p,q,k}(n)$ = the weight from neuron p to neuron q, at n^{th} step, where q lies in the layer k and neuron p in $(k-1)^{th}$ layer counted from the input layer;

$\delta_{p,k}$ = the error generated at neuron q, lying in layer k;

$Out_{p,j}$ = output of neuron p, position of layer j.

For generating error at neuron p, lying in layer j, we use the following expression[2]:

$$\delta_{p,j} = Out_{p,j}(1-Out_{p,j})(\sum_{q} \delta_{q,k}.w_{p,q,k}) \quad \text{..... (4.4)}$$

where:

$$q \varepsilon \{q_1, q_2, q_3\}$$

For training a network by this algorithm, one has to execute the following 4 steps in order for all patterns one by one.

For each input-output pattern do begin

1. Compute the output at the last layer through forward calculation; 2. Compute δs at the last layer and propagate it to the previous layer bu using expression (4.4); 3.Adjust weights of each neuron by using expression (4.2) and (4.3) in order; 4.Repeat from step 1 until the error at the last layer is within a desired margin.

End For;

5. KNOWLEDGE UPDATING PROCESS

In this system, Machine learning is knowledge based system which is updated from external input, and the steps are:

1. The initial condition of machine has no information yet (empty data); 2. External input is person name which is used initial knowledge. Is the name classified to a man or woman gender?; 3.For each given gender suggestion, the system will accept the external input as an addition knowledge to the system.

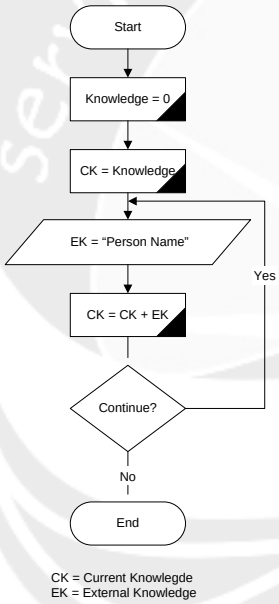


Figure 2. Knowledge updating process

6. SYLLABLE PATTERN RECOGNITION PROCESS

In each specific area, every parent will give the name of the child based on the customs and cultures in the area. The name given by parents has a pattern based on a word. The processes of pattern recognition that a word will be done in this study are:

1. Words separation; 2.Word syllable separation; 3.Search words and syllables in the database; 4.Retrieval of gender information for each syllable found; 5.Giving weight value to each syllable based on the amount of data found; 6.Calculation of weighted average of each syllable; 7.Gender information advisory. 8.Gender information verification as a new external input (knowledge updating).

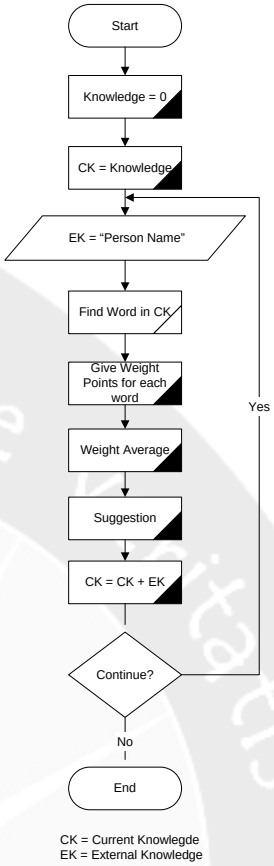


Figure 3. Syllable pattern recognition process

7. CONCLUSION

The conclusions are:

Provision of gender information can be made based on the name of people. In general, the naming can be based on the geographic location. .Giving weight value of each syllable is based on the number of syllables discovery in the data previously saved. Weighted average gives advice based on gender information included names. Gender information can be used as one factor for the provision of attributes or subsequent decisions.

8. REFERENCES

[1] Building on Gender, Agrobiodiversity and Local Knowledge". FAO, 2004.

[2] Konar, Amit. Artificial intelligence and soft computing: behavioral and cognitive modeling of the human brain. 2000

[3] Xinyong, Qiao, Liu Wei, A Method for Evaluating the Sensitivity of Signal Features in Pattern Recognition Based on Neural Network, Department of Mechanical Engineering Academy of Armored Forces Engineering Beijing, China, 2008

Maintaining Visibility of A Moving Target: The Case of An Adaptive Collision Risk Function

Ashraf Elnagar
Department of Computer science
University of Sharjah
Sharjah, UAE
ashraf@sharjah.ac.ae

Ibrahim Al-Bluwi
LAAS-CNRS
Universitt'e de Toulouse
F-31077, Toulouse, France
al.bluwi@laas.fr

ABSTRACT
This paper presents a novel approach for the problem of tracking a moving target in a dynamic environment. The robot has to move such that it keeps the target visible for the longest time possible while avoiding collision with moving obstacles. The solution consists of three interacting components which perform: tracking, collision avoidance and motion selection. For collision avoidance, an adaptive collision risk value is computed. This risk represents the likelihood of the pursuer colliding with any of the moving obstacles using the occluding and collision risks, the decision maker decides on the next safe move. Experimental results confirmed the robustness of the proposed algorithm for tracking in a dynamic environment amidst obstacles.

Keywords
Target tracking, dynamic environments, collision avoidance.

1. INTRODUCTION
A recent study estimated the number of robots world-wide that were in operation in 2007 to be close to 6.5 million [4]. Today, the use of robots is not restricted to factories and highly structured environments any more. Robots today perform daily-life tasks like, vacuum cleaning and grass mowing. They also guide people in museums and hospitals assist the elderly and blind and work as security watchmen. In fact, applications of robots are not restricted to the aforementioned, and are continuously increasing in number and type. This expansion puts today's robots against two main challenges. Namely, the ability to perform more complex tasks and to work in unpredictable and unknown environments. This paper considers the task of target following, which is a natural task performed by humans. It also assumes that moving objects exist around the robot performing the tracking task, which reflects the nature of our daily-life environments. Having robots with such a capability opens the door for many applications in several domains such as medicine, security, and home entertainment.

The paper is organized as follows: Section 2 surveys previous work on tracking in dynamic environments. The problem is then further described and formalized in Section 3. Collision advisor and decision maker components are explained in Sections 4 and 5, respectively. Simulations and results are discussed in Section 6 and the paper then ends with a section for conclusions and possible future work.

The importance of this study is of two fold. The first is that it addresses a lively problem that has applications in various domains and the second is that it contributes to filling the gap in current research on the problem of autonomous tracking. Current research assumes that tracking occurs in a static environment, where the target is the only moving object in the environment. This assumption allows concentrating on finding a solution for the tracking problem. However, it is very restricting as most application domains are in dynamic environments.

The proposed solution in this paper is based on an architecture that is made of three distinct but interacting components. These components are called the occlusion advisor, collision advisor and decision maker. The occlusion advisor is dedicated for tracking, the collision advisor for collision avoidance and the decision maker for motion selection. The robot moves in the environments trying to minimize two risks: the risk of losing the target and the risk of colliding with a moving object

2. PREVIOUS WORK
Research related to the problem of tracking in dynamic environments can be categorized into two groups: work done on tracking in static environments and work done on motion planning in dynamic environments. The following is a survey of work in these groups.

The problem of tracking in a static environment is formally defined and analyzed using Game Theory, [13]. In other works, such as [15], the problem is modeled as a motion planning problem of a rod, of variable length. It was shown that the problem of tracking a target around one corner is completely decidable, [2]. However, the decidability of the general problem in a cluttered environment was found to be at least NP complete, [16]. In [3], the environment is divided into decidable and undecidable regions and heuristics were used to approximate the bounds of the decidable regions.

All of the above mentioned results are of theoretical significance for the problem of autonomous tracking. However, practical solutions are also reported. In [7] a tracking algorithm is introduced that maximizes the target's shortest distance to escape. Similarly, [1] uses a greedy local strategy that is based on a different risk function, which tries to achieve balance between not losing the target while keeping the target visible for the longest time.

A plethora of algorithms have been introduced to address the problem in dynamic environments. For obstacles with a predictable motion track, several extensions of available planning algorithms in static environments are proposed. For example, the joint time-configuration space idea, [8], and the velocity decomposition technique, [11].

Online algorithms also exist which assume that the motions of the obstacles are not known in advance. Some adopt a plan-then-fix strategy, like the D* [12] and the DRRT [6] which repeatedly observes the environment and fixes an initial RRT plan by adding and removing nodes. A similar approach is used with the PRM [9]. Other algorithms adopt a plan-then-improve strategy. A rough plan is initially created and then is improved over time to adopt for the changes in the environment, [17]. Such planners are called anytime planners since they are ready to give a plan any time.

Several algorithms for motion prediction using different techniques are reported. For example, Kalman filters [10], Brownian models [14] and auto-regressive models [5] have been proposed to produce predictions based on an observed short history of the obstacle's movements. Other approaches also follow a learn-then-predict approach [18]. In such approaches, the environment is first observed and motion patterns are recorded online. A model is then created based on the observed data and is used to produce predictions on-line. In other works, such as [18], the motion patterns are clustered and prediction is done by mapping the newly observed motion patterns to the closest cluster. The representative of this cluster is then used to produce a prediction. Some other techniques (e.g., [19]) learn motion patterns on-line. However, the learn-then-predict model is not always suitable since it restricts the application to previously known environments. It follows the intuition that objects usually do not move randomly, but seek to reach goals.

Previous work has addressed the two distinct problems: tracking in static environments and motion planning in dynamic environments separately. This paper studies the effect of addressing both problems together and how the performance of the robot is affected while tracking in a global dynamic environment amidst randomly moving obstacles.

3. PROBLEM DESCRIPTION

We use the terms pursuer and evader to refer to the robot and the target respectively. Both, the pursuer and the evader are assumed to be rigid objects that are not restricted by any constraints other than having the same velocity bound. The evader is assumed to move at random in the environment and the pursuer has the ability to detect the evader and the dynamic obstacles.

Let p and e denote the pursuer and evader respectively. Without loss of generality, we assume p and e to be points moving in the configuration space. We refer to a time step in the tracking problem with , where $0 \leq k \leq T$. The current state of the pursuer at time is denoted by . It belongs to the state space of the pursuer χ^p . Similarly, the current state of the evader is denoted by and it belongs to χ^e; the evader's state space.

At each time step, the pursuer and evader choose an action and from their respective action spaces and . Tracking takes place in the Euclidean workspace W . This workspace is cluttered with a set of static obstacles which occupy the space , and the free part of W is referred to as . Thus, $W = W_{obs} \cup$. The set of dynamic obstacles in W is referred to as O. At each time step, we are interested only in dynamic obstacles that are currently visible to the pursuer (O_k).

We say a point q is visible to p if and only if it can be connected to it with a straight line l. Two constraints govern l. First, it should lie completely inside . Second, the length of l should be less than ρ, where ρ is the maximum reachable distance by the pursuer's sensors. We are now able to define $V \subseteq$ as the visibility region of p. Any point can be connected to p with l. Note that our definition of prohibits the intersection of l with obstacles in and not in O. This means that we assume in our problem that dynamic obstacles do not cause occlusion. Another important note is that when tracking starts, we assume that the evader is initially detected, i.e. . It is now possible to define our tracking problem as a decision problem over the triplet $\{\Omega, D, C\}$, where:

- Ω is the problem state space. It represents all possible combinations of evader and pursuer states with possible states of dynamic obstacles. At time t_k the problem state is denoted as ω_k and it captures , and the states of each obstacle in O at that time instance.
- D is the set of possible decisions. A decision d_k made at time is a choice of action that the pursuer makes after observing Ω.
- C is the set of possible consequences. A consequence $c(d_k)$ is a result or an outcome that occurs when the pursuer makes the decision .

Depending on the goal which the pursuer would like to achieve, it will have a preference for some consequences over others. Based on this preference, the pursuer should choose its decisions from D at each time step. More specifically, each consequence c has a utility value that quantifies to what extent the pursuer likes the consequence. This utility is defined as the function U over the decision that caused this consequence. We use the notation to denote preference between decisions and pronounce as is preferred to . We say $d_k \geq d'_k$ only if $U(d_k) \geqslant U$.

The goal of the pursuer in our tracking problem now becomes to choose at each time step such that:

$$d_k = \underset{d \in D}{\operatorname{argmax}} \{U(d)\}$$

Our tracking problem is now clearly dependent on how we define the utility function U. Roughly speaking, U should award more decisions that make the pursuer less susceptible to loosing the evader and less susceptible to colliding with a dynamic obstacle at the same time. In the following, we give two general definitions for U depending on the tracking goal which the pursuer aims at.

4. THE COLLISION ADVISOR

Our solution is intuitively divided into three separate components. The first component helps the pursuer evaluate the tracking status. The second helps to evaluate the safety status, and the third component helps to make a motion decision (i.e., where to move next). We refer to these components as the "occlusion advisor", the "collision advisor" and the "decision maker" respectively.

At each time step, the robot will observe the current state of the environment and two risk values are produced. The first is by the occlusion advisor and it represents the likelihood of losing the evader and the second is by the collision advisor and it represents the likelihood of colliding with a moving obstacle. The decision maker then uses these two risk values to choose a control action that is best in terms of tracking effectiveness and robot safety. Next, we discuss in detail the collision advisor component.

When dealing with unknown dynamic obstacles, it is very useful to be able to predict the future movements of these obstacles. This allows planners to create plans that are safer and less likely to change. In fact, motion prediction is a natural task that we perform in our daily life as we move around and avoid colliding with other moving entities like people and cars. In the collision risk function to be introduced in the next section, we assume that a motion prediction algorithm exists and is used by the pursuer.

4.1 Risk Computation Algorithm

In a general environment, occasionally the pursuer is expected to find itself among several dynamic obstacles moving in arbitrary directions. Notice that such dynamic obstacles are assumed to be non-cooperative. The risk function to be introduced in this section will help the pursuer measure the risk of colliding with any of the moving obstacles in similar scenarios.

Assume that a robot is placed at some room and due to some unknown error it starts moving around itself randomly. To be safe from being hit by this robot, the intuitive behavior is to keep away from it. The farther, the better, and the direction does not matter since it is not known where the robot will move next. Consider on the other hand the case of crossing a road. Cars move in one direction and to be safe from being hit by a moving car, one has to avoid being in front of it. Therefore, it is usually safe to get close to the car from its sides and its rear. These two scenarios depict the difference between the case of predictable and unpredictable obstacle motions. In the first case, it is not possible to predict where the malfunctioning robot is going to move next, so it is required to maintain a clearance distance from the robot in all directions. However, it is possible in the second case to predict where the car will move next, so a clearance distance can be maintained based only on the direction of the predicted motion. Hence we distinguish between two risk types: random risk and prediction risk.

Random risk is independent of the motion direction of the obstacle. It represents the likelihood of colliding with an obstacle regardless of where this obstacle is moving. Such a risk can be represented as follows:

$$Random\ Risk = 1 - \frac{d^m}{d^m_{max}}$$

where d is the distance to the obstacle, d is the maximum distance possible to the obstacle and m is a scaling factor. This equation creates a normalized risk value, where a closer distance to the obstacle yields a higher risk value.

Prediction risk, on the other hand, is dependent on where the observer anticipates the obstacle will move next. If the observer knows for sure the very location of the next move, then the risk will be *1* at that location and *0* everywhere else. However, it is usually not possible to predict the exact location of the next move. There is always a degree of uncertainty based on the amount of information available and the prediction method used.

We model prediction risk as a circle around the prediction point, where the radius of this circle depends on the confidence in this prediction. If the prediction is guaranteed to be correct then the radius is *0*, otherwise the radius of the circle will grow proportional to the lack of confidence. Therefore, prediction risk can be modeled using the following equation:

$$PredictionRisk = (1 - confidence) * (1 - \frac{dP^m}{dP^m_{max}})$$

where dP^m is the distance to the prediction point, dP^m_{max} is the maximum distance possible to the prediction point, and confidence is a value between *0* and *1* that represents our belief in the validity of the prediction.

4.2 An Adaptive Risk Function for Collision

Generally speaking, both random and prediction risks should be considered when avoiding dynamic obstacles. In our malfunctioning robot example, both risks exist, but prediction risk is dominated by random risk since we have no confidence in our predictions. In the car example, prediction risk is dominant, however, random risk should not be completely neglected as the car may for some reason choose to reverse its motion or take a sharp turn. These two cases represent two extremes where one risk dominates the other. In the normal case where motion is not completely random and not completely structured, both random and prediction risks should have a say.

We represent our collision risk function for one dynamic obstacle as a weighted sum of these two risks, where increasing the weight of one risk causes a decrease in the weight of the other.

$$\varphi = (1 - \lambda)(RandomRisk) + \lambda\ PredictionRisk$$

The weight factor λ may be interpreted as the prediction confidence. The more the observer is doubtful about its predictions the more significant the random risk should be and vice versa. When the observer has no confidence at all in the predictions, only random risk will be functioning, whereas when the observer is totally confident of its predictions only the prediction risk will be functioning. For the other cases, the overall risk will be a balance between the two. The value is an aggregation of the individual values for each of the dynamic obstacles. A pessimistic risk computation like the one used for aggregating the risks of the occluding vertices may also be used.

4.3 Confidence, Error Rates and Adaptation

What remains in the definition of the risk function is to identify how confidence is computed. It is intuitive to think of confidence as a function of the errors reported by the prediction model. The confidence factor is inversely proportional to the amount of prediction errors. For this reason, we introduce the two terms prediction error and prediction error rate.

The prediction error measures how bad a single prediction is, whereas the prediction error rate measures how bad a prediction model performs in general. If *dist()* is a function that measures the distance between two points, then prediction error can be computed as follows:

$$\begin{cases} 1 & if\ dist(pred, actual) > dist(obs, actual) \\ 0 & if\ dist(pred, actual) = 0 \\ \frac{dist(pred, actual)}{dist(orig, actual)} & otherwise \end{cases}$$

where *pred* is the predicted location of the obstacle, actual is the location to which it has moved and *orig* is the original location of the obstacle. This function penalizes predictions that are far from the actual location to which the obstacle has moved. The penalty is relative to the distance actually traveled by the obstacle.

The error rate(er) can now be computed by taking the average of all of the prediction errors computed for predictions done in a certain environment. The error rate can be different for the same prediction model over different environments. This is because the model can succeed at predicting the motion of certain types of obstacles and fail at predicting others. Therefore, the error rate should be computed for each environment separately. Confidence now becomes the inverse of the error rate and the collision risk function for one dynamic obstacle becomes as follows:

$$\varphi = er(RandomRisk) + (1 - er)PredictionRisk$$

$$\varphi = er\left(1 - \frac{d^m}{vr^m}\right) + (1 - er)\left(1 - \frac{dP^m}{vr^m}\right)$$

The error rate can also be computed incrementally. The observer begins moving with an error rate of 1, which causes the collision risk to be completely dependent on the random risk part of the function. As the observer moves and makes predictions, the error rate is updated and more weight is given to the prediction risk according to the success made at these predictions.

Such incremental computation of the error rate allows the observer to adapt to changes in the behavior of the obstacles in the environment. For the example, consider the case when a set of obstacles moves in a predictable manner for some time and then start moving in an unpredictable way, the incremental computation of the error rate will allow the observer to rely on prediction risk more at the beginning and on random risk more later on.

This incremental computation also gives room for the application of learning algorithms. If a learning algorithm is used to learn the motion patterns of obstacles in the environment and this algorithm enhances the prediction model over time, the incremental computation will allow the observer to rely more on the prediction risk as the prediction algorithm performs better over time.

Figure 1 illustrates the concept of random and prediction risks and shows how they change over time with the change of the error rate. In (a), the error rate is very high so random risk is dominant. As we move through (b), (c) and (d), the prediction risk becomes more visible as the error rate drops. More illustrative examples are provided in the next section.

5. DECISION MAKING

The goal of the pursuer is to track in a dynamic environment. The goal of the pursuer is to choose the decision that has the maximum utility possible, U. The utility of a decision depends on the consequence produced by this decision. A consequence in our case is basically a new pursuer configuration. The worst consequences possible are those that prevent the pursuer from seeing the evader or those that cause collision with a dynamic obstacle. Decisions causing such consequences should have a utility value of 0. The utility of other decisions is based on the collision and occlusion risks caused by these decisions. For the case of maximizing escape time (U), it can be redefined as:

$$U(d_k) = \begin{cases} 0 & if\ e \in V \\ 0 & if\ p \in O \cup W_{obs} \\ 1 & otherwise \end{cases}$$

where *overallRisk* is a normalized aggregation of the two main risks: occlusion risk ( ) and collision risk ( ).

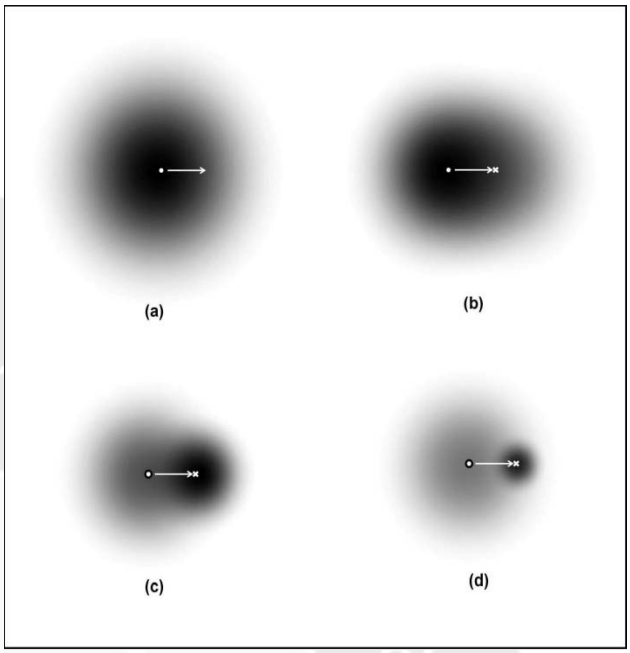


Figure 1. An illustration of the adaptive collision risk function. The darker regions represent higher risk and the predicted location is marked with a cross.

6. SIMULATION RESULTS

In this section, we show some simulation examples for the use of the introduced risk function in computing collision risk. The application used for simulation allows the user to choose a location for the observer and the number of dynamic obstacles. The application then moves the obstacles in the environment and performs two tasks. The first is a measurement for at the location of the observer, where this measurement is printed numerically along with the error rate to the console. The second is a computation for at each point in the environment. The computed values are not printed to the console but are given a shade of red to represent the risk value. Lighter red means less risk and vice versa.

For simulation purposes, we used a simple prediction model. This model considers only the previous three moves of the obstacle, computes the acceleration and estimates the next move accordingly. The following are several screen shots for several runs of the simulator with different error rates and distributions of the dynamic obstacles. The computed collision risks and error rates are shown in the figure captions.

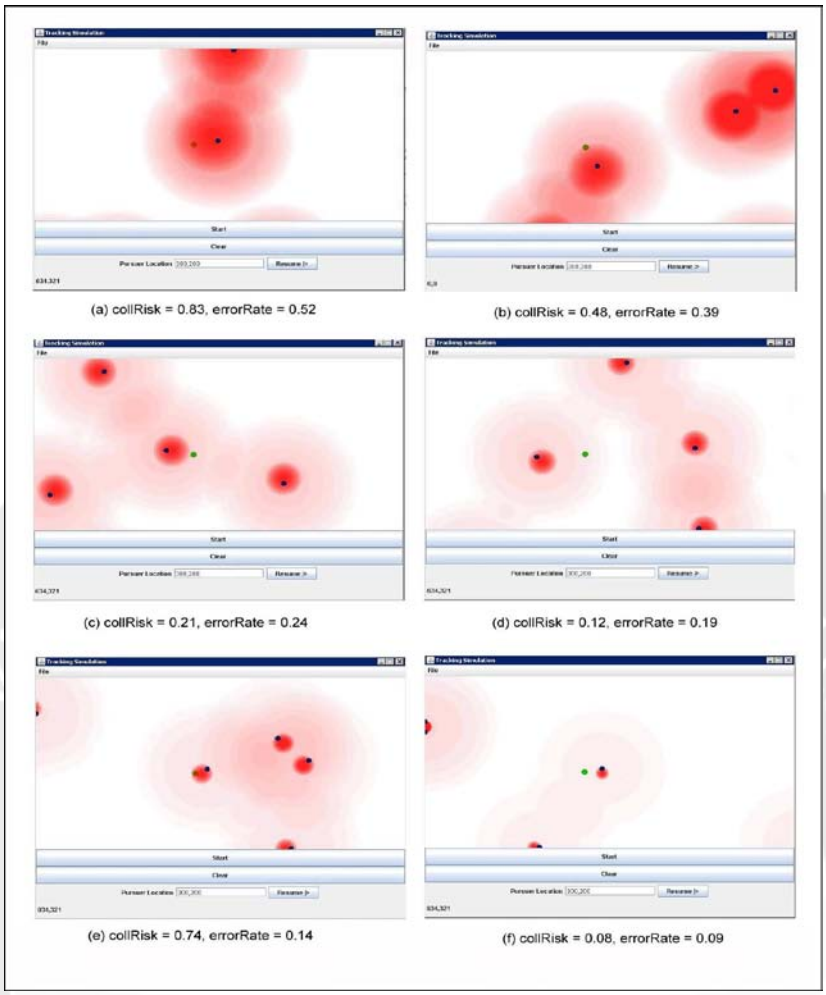


Figure 2. The shape of the collision risk function in six different scenarios.

- **Corners in the environment**: the higher the number of corners is, the harder tracking becomes. This is because the evader has more escape choices and the pursuer has to take them all in mind.
- **Density**: it is the percentage of the environment that is occupied by obstacles. The higher the density is, the harder tracking is. This is because the pursuer has less motion choices in a dense environment.
- **Dynamic obstacles**: the higher the number is, the harder tracking becomes. This is because the pursuer has to perform more collision avoidance motions, which may negatively affect the tracking process.

We have performed several simulations to investigate the impact of the above parameters on the tracking process in three different environments. Namely, the first is with high density and few corners, the second is with moderate density and corners, and the last one is with high density and corners. For each of these environments, the simulations were conducted repeatedly after adding 0, 10, 20, and 30 randomly-moving obstacles.

Table 1 summarizes the results of the 12 simulation sets. The table shows the pursuer's reaction in the dynamic environments. Each dynamic simulation is carried out 30 times. For each run, the table reflects the number of collisions between the pursuer and the dynamic obstacles on average in the 30 runs. The table also shows the percent of time, on average, for which the evader remained visible to the pursuer in the 30 runs. The results confirmed the validity and robustness of the proposed system.

Table 1. Summary of simulation results

Averages	Scenario	No Dyn. Obs.	10 Dyn. Obs.	20 Dyn. Obs.	30 Dyn. Obs.
% of Time Visible	First	97.8%	89.5%	89.5%	86.6%
Number of Collisions		0	0.6	2.6	5.1
% of Time Visible	Second	100%	96.5%	92.9%	90.6%
Number of Collisions		0	1	1.6	2.9
% of Time Visible	Third	99.6%	88.2%	74.3%	54.4%
Number of Collisions		0	0.86	1.2	3.8

But why can't the pursuer avoid all collisions despite of having an adaptive collision avoidance advisor? In many simulation runs, the pursuer succeeds to complete the tracking task without any collision. However, in other runs, it may collide with some dynamic obstacles. This is attributed to either of the following reasons. The first one is the fact that any dynamic obstacle moves in a straight line until it collide with a static obstacle or environment boundary, then it will bounce in a new random direction. This abrupt motion may lead to a collision if the pursuer is close by. Second, the pursuer may be trapped among several dynamic obstacles to the extent that collision becomes inevitable even if motion prediction is correct. This is because dynamic

obstacles are assumed to be non-cooperative and, therefore, such obstacles may head to the pursuer. The third reason is due to the degree of trade-off between safety and tracking. This is inherent in the definition of the overall risk function which assigns weights for the collision and occlusion risks. In reality, sometimes it is acceptable to allow collisions during tracking. This depends on the type of pursuer used (fragile, flexible, solid, etc.), type of dynamic obstacles present (soft, hard, etc.), and type of application in hand. It is fair to compare our robot pursuer with a human pursuer. A human pursuer tracking in a cluttered environment like a show room is also susceptible to collisions with objects/humans in the environment. The criticality of these collisions relies on the situation and the type of people/objects with which collisions occur. In fact, it is almost impossible to completely avoid collisions and, therefore, the goal should always be to minimize these collisions as much as possible.

7. CONCLUSIONS

In this paper, a novel approach for the problem of tracking a moving target in a dynamic environment is presented. The solution has three interacting components: tracking, collision avoidance and motion selection. The collision advisor produces an adaptive collision risk value which guides the pursuer to avoid colliding with the moving obstacles present in the environment. To compute this risk factor, , visible dynamic obstacles, to the pursuer, along with a short history of their random motion track are made available. This risk represents the likelihood of the pursuer colliding with any of the moving obstacles. Based on this risk and the occlusion risk, the decision maker component evaluates the criticality of the tracking situation. The proposed solution is validated using a comprehensive set of simulations, which show that transition from tracking in static environments to tracking in dynamic environments can be done without much loss in robot safety or tracking ability.

8. REFERENCES

[1] T. Bandyopadhyay, Y. Li, M. Ang Jr, and D. Hsu. A Greedy Strategy for Tracking a Locally Predictable Target among Obstacles. In Proceedings of the IEEE International Conference on Robotics and Automation, pages 2342–2347, 2006.

[2] S. Bhattacharya, S. Candido, and S. Hutchinson. Motion Strategies for Surveillance. In Proceedings of Robotics: Science and Systems III, pages 249–256, Atlanta, GA, USA, June 2007. MIT Press.

[3] S. Bhattacharya and S. Hutchinson. Approximation Schemes for Two-Player Pursuit Evasion Games with Visibility Constraints. In Proceedings of Robotics: Science and Systems IV, Zurich, Switzerland, June 2008. MIT Press.

[4] T. I. S. Department. 2007: 6,5 million robots in operation world-wide, October 2008. Retrieved Feb. 2009 from the world wide web: http://www.worldrobotics.org.

[5] A. Elnagar and A. Hussein. An adaptive motion prediction model for trajectory planner systems. In Proceedings of the IEEE International Conference on Robotics and Automation, volume 2, pages 2442–2447, 2003.

[6] D. Ferguson, N. Kalra, and A. Stentz. Replanning with RRTs. In Proceedings of the IEEE International Conference on Robotics and Automation, pages 1243–1248, 2006.

[7] H. Gonzalez-Banos, C. Lee, and J. Latombe. Real-time combinatorial tracking of a target moving unpredictably among obstacles. In Proceedings of the IEEE International Conference on Robotics and Automation, volume 2, pages 1683–1690, 2002.

[8] D. Hsu, R. Kindel, J. Latombe, and S. Rock. Randomized Kinodynamic Motion Planning with Moving Obstacles. The International Journal of Robotics Research, 21(3):233–255, 2002.

[9] L. Jaillet and T. Simeon. A prm-based motion planner for dynamically changing environments. Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, 2:1606–1611, 2004.

[10] R. Kalman. A new approach to linear filtering and prediction problems. Journal of Basic Engineering, 82(1):35–45, 1960.

[11] K. Kant and S. Zucker. Toward Efficient Trajectory Planning: The Path-Velocity Decomposition. The International Journal of Robotics Research, 5(3):72–89, 1986.

[12] S. Koenig and M. Likhachev. D* lite. In Proceedings of the Eighteenth national conference on Artificial intelligence, pages 476–483. American Association for Artificial Intelligence, 2002.

[13] S. LaValle, H. Gonzalez-Banos, C. Becker, and J. Latombe. Motion strategies for maintaining visibility of a moving target. In Proceedings of IEEE International Conference on Robotics and Automation, volume 1, pages 731–736, 1997.

[14] M. Montemerlo, S. Thrun, and W. Whittaker. Conditional particle filters for simultaneous mobile robot localization and people-tracking. In Proceedings of the IEEE International Conference on Robotics and Automation, volume 1, pages 695–701, 2002.

[15] R. Murrieta, A. Sarmiento, and S. Bhattacharya, S. Hutchinson. Maintaining visibility of a moving target at a fixed distance: the case of observer bounded speed. In Proceedings of the IEEE International Conference on Robotics and Automation, volume 1, pages 479–484, 2004.

[16] R. Murrieta-Cid, R. Monroy, S. Hutchinson, and J. Laumond. A Complexity result for the pursuit-evasion game of maintaining visibility of a moving evader. In Proceedings of IEEE International Conference on Robotics and Automation, pages 2657–2664, 2008.

[17] J. van den Berg, D. Ferguson, and J. Kuffner. Anytime path planning and replanning in dynamic environments. In Proceedings of the IEEE international Conference on Robotics and Automation, pages 2366–2371, 2006.

[18] D. Vasquez and T. Fraichard. Motion prediction for moving objects: a statistical approach. Proceedings of the IEEE international Conference on Robotics and Automation, 4:3931–3936, 2004.

[19] D. Vasquez Govea, T. Fraichard, and C. Laugier. Incremental learning of statistical motion patterns with growing hidden markov models. In Proceedings of the International Symposium of Robotics Research, Hiroshima, Japan, 2007.

Measuring Interesting Rules in Characteristic Rule

Spits Warnars
Department of Computing and Mathematics, Manchester Metropolitan University
John Dalton Building, Chester Street
Manchester M1 5GD, United Kingdom
+44 (0)161 247 1779
s.warnars@mmu.ac.uk

ABSTRACT
Finding interesting rule in the sixth strategy step about threshold control on generalized relations in attribute oriented induction, there is possibility to select candidate attribute for further generalization and merging of identical tuples until the number of tuples is no greater than the threshold value, as implemented in basic attribute oriented induction algorithm. At this strategy step there is possibility the number of tuples in final generalization result still greater than threshold value. In order to get the final generalization result which only small number of tuples and can be easy to transfer into simple logical formula, the seventh strategy step about rule transformation is evolved where there will be simplification by unioning or grouping the identical attribute. Our approach to measure interesting rule is opposite with heuristic measurement approach by Fudger and Hamilton where the more complex concept hierarchies, more interesting results are likely to be found, but our approach the simpler concept hierarchies, more interesting results are likely to be found and the more complex concept hierarchies, more complex process generalization in concept tree. The decision to find interesting rule is influenced with wide or length and depth or level of concept tree.

Keywords
Attribute oriented induction, Concept tree, Heuristic Measurement.

1. INTRODUCTION
Attribute oriented induction approach is developed for learning different kinds of knowledge rules such as characteristic rules, discrimination or classification rules, quantitative rules, data evolution regularities [1], qualitative rules [2], association rules and cluster description rules [3]. Attribute oriented induction has concept hierarchy as an advantage where concept hierarchy as a background knowledge which can be provided by knowledge engineers or domain experts [3-5]. Concepts are ordered in a concept hierarchy by levels from specific or low level concepts into general or higher level and generalization is achieved by ascending to the next higher level concepts along the paths of concept hierarchy [8].
DBLearn is a prototype data mining system which developed in Simon Fraser University integrates machine learning methodologies with database technologies and efficiently and effectively extracts characteristic and discriminant rules from relational databases [9,10]. Since 1993 DBLearn have led to a new generation of the system call DBMiner with the following features:

a. Incorporating several data mining techniques like attribute oriented induction, statistical analysis, progressive deepening for mining multiple-level rules and meta-rule guided knowledge mining [11] data cube and OLAP technology [12].
b. Mining new kinds of rules from large databases include multiple level association rules, classification rules, cluster description rules and prediction.
c. Automatic generation of numeric hierarchies and refinement of concept hierarchies.
d. High level SQL-like and graphical data mining interfaces.
e. Client server architecture and performance improvements for larger application.
f. SQL-like data mining query language DMQL and Graphical user interfaces have been enhanced for interactive knowledge mining.
g. Perform roll-up and drill-down at multiple concept levels with multiple dimensional data cubes.

DBMiner had been developed by integrating database, OLAP and data mining technologies[12] which previously called DBLearn have their own database architecture. Concept hierarchy is stored as a relation in the database provides essential background knowledge for data generalization and multiple level data mining. Concept hierarchy can be specified based on the relationship among database attributes or by set groupings and be stored in the form of relations in the same database [11]. Concept hierarchy can be adjusted dynamically based on the distribution of the set of data relevant to the data mining task and hierarchies for numerical attributes can be constructed automatically based on data distribution analysis [11].

For making easy the implementation a concept hierarchy will just only based on non rule based concept hierarchy and just learning for characteristic rule. Characteristic rule is an assertion which characterizes the concepts which satisfied by all of the data stored in database. Provide generalized concepts about a property which can help people recognize the common features of the data in a class. For example the symptom of the specific disease [6].
For doing the generalization there are 8 strategy steps must be done [4], where step 1 until 7 as for characteristic rule and step 1 until 8 for classification/discriminant rule.

a. Generalization on the smallest decomposable components
b. Attribute removal
c. Concept tree Ascension
d. Vote propagation
e. Threshold control on each attribute
f. Threshold control on generalized relations
g. Rule transformation
h. Handling overlapping tuples

2. PROBLEM IDENTIFICATION

In the sixth strategy step about threshold control on generalized relations, there is possibility to select candidate attribute for further generalization and merging of identical tuples until the number of tuples is no greater than the threshold value, as implemented in basic attribute oriented induction algorithm [4]. At this strategy step there is possibility the number of tuples in final generalization result still greater than threshold value. In order to get the final generalization result which only small number of tuples and can be easy to transfer into simple logical formula, the seventh strategy step about rule transformation is evolved [4] where there will be simplification by unioning or grouping the identical attribute [2,4]. Based on the above explanation then there are problems like :

a. Which one the best attribute for further generalization ?

b. Which one the best attribute for further simplification ?

Our implementation attribute oriented induction characteristic rule has been implemented with Java programming language and MySQL database with 50.000 records, while data example and concept hierarchy refer to [4,6]. Based on concept hierarchy in [4,6] we have 4 concept trees, they are :

a. Figure 1 is concept tree for major.

b. Figure 2 is concept tree for category.

c. Figure 3 is concept tree for birthplace

d. Figure 4 is concept tree for GPA.

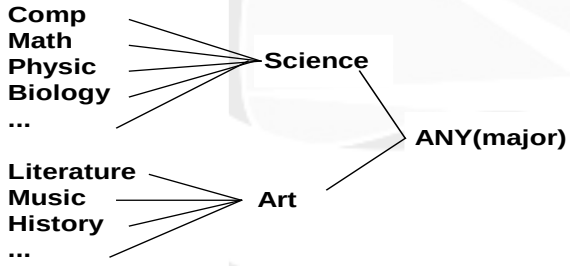


Figure 1. Concept tree for major

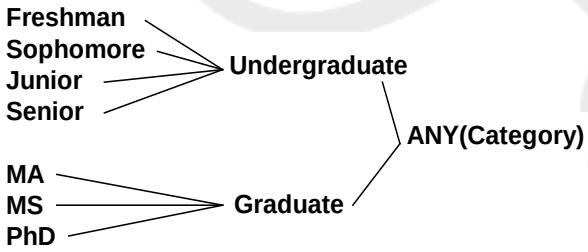


Figure 2. Concept tree for category

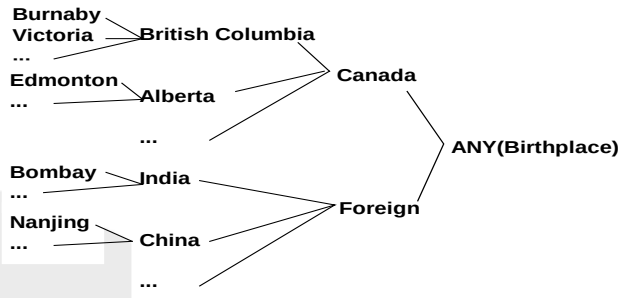


Figure 3. Concept tree for birthplace

Figure 4. Concept tree for GPA

Figure 5 show the result when program was run to find characteristic rule for graduate student with threshold 2 and stop after the fifth strategy step about threshold control on each attribute.

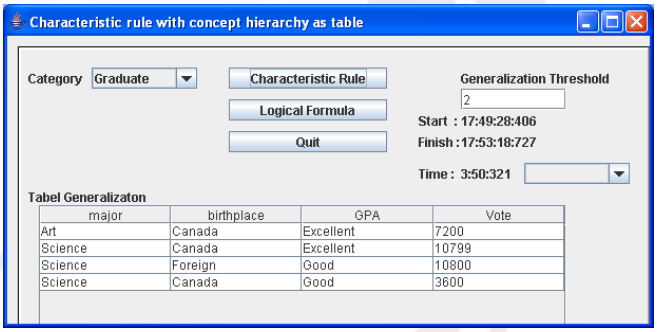


Figure 5. Result for threshold=2 after threshold control on each attribute

Based on generalization steps for characteristic rule, when the number of distinct tuples still greater than threshold control then the next strategy step which is the sixth strategy step must be done[4]. At explained before because there is possibility the number of tuples still greater then the seventh strategy step must be done. Table 1 until 6 show the possibilities the final generalization include with the rules.

Table 1. Further generalization on major attribute and unioning on birthplace attribute

Major	Birthplace	GPA	Vote
ANY	Canada	{Excellent, Good}	21599
ANY	Foreign	Good	10800

birthplace(x) ∈ Canada^GPA(x)∈{Excellent,Good}[66.66%] V

birthplace(x) ∈ Foreign ^ GPA(x) ∈ Good [33.33%]

Table 2. Further generalization on major attribute and unioning on GPA attribute

Major	Birthplace	GPA	Vote
ANY	Canada	Excellent	17999
ANY	{Foreign,Canada}	Good	14400

birthplace(x) ∈ Canada ^ GPA(x) ∈ Excellent [55.55%] V

GPA(x) ∈ Good [44.44%]

Table 3. Further generalization on birthplace attribute and unioning on major attribute

Major	Birthplace	GPA	Vote
{Art,Science}	ANY	Excellent	17999
Science	ANY	Good	14400

GPA(x) ∈ Excellent [55.55%] V

major(x) ∈ Science ^ GPA(x) ∈ Good [44.44%]

Table 4. Further generalization on birthplace attribute and unioning on GPA attribute

Major	Birthplace	GPA	Vote
Art	ANY	Excellent	7200
Science	ANY	{Excellent,Good}	25199

major(x) ∈ Art ^ GPA(x) ∈ Excellent [22.22%] V

major(x) ∈ Science ^ GPA(x) ∈ {Excellent, Good} [77.77%]

Table 5. Further generalization on GPA attribute and unioning on major attribute

Major	Birthplace	GPA	Vote
Art	Canada	ANY	7200
Science	{Canada, Foreign}	ANY	25199

major(x) ∈ Art ^ birthplace(x) ∈ Canada [22.22%] V

major(x) ∈ Science [77.77%]

Table 6. Further generalization on GPA attribute and unioning on birthplace attribute

Major	Birthplace	GPA	Vote
{Art,Science}	Canada	ANY	21599
Science	Foreign	ANY	10800

birthplace(x) ∈ Canada [66.66%] V

major(x) ∈ Science ^ birthplace(x) ∈ Foreign [33.33%]

3. DEPTH AND LENGTH OF CONCEPT TREE

The final generalization results in table 1 into 6 have the equal interesting rule, the same important and the best result will depend on user's interest. In order to find the best final generalization from six final generalization results in table 1 into 6, where automatically can be built by program application. Our approach to measure interesting rule is influenced by heuristic measurement approach by Fudger and Hamilton where the more complex concept hierarchies, more interesting results are likely to be found [7]. Opposite with Fudger and Hamilton approach, in our approach the interesting rule can be found in the simple concept hierarhies. The simpler concept hierarchies, more interesting results are likely to be found and the more complex concept hierarchies, more complex process generalization in concept tree. The decision to find interesting rule is influenced with wide or length and depth or level of concept tree:

a. Depth or level of concept tree, where simple depth or level in concept tree will have simple generalization process in concept tree, but the more depth or level in concept tree will have more generalization process in concept tree.
b. Wide or length of concept tree or amount of concepts per level in concept tree, where simple concepts will have simple generalization process in concept tree, but the more concepts will have more generalization process in concept tree.

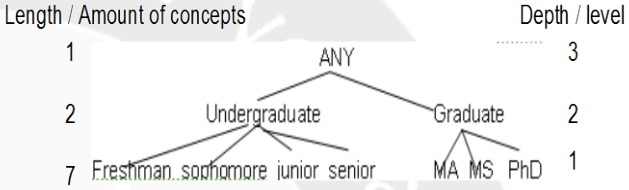


Figure 6. Depth and length of category concept tree

For example, figure 6 shows category concept tree has 3 levels and each of level has wide or length of concepts where level 3 as the highest level must have 1 concept, the next level 2 has 2 concepts are undergraduate and graduate and the last level 3 has 7 concepts are Freshman, Sophomore, Junior, Senior, MA, MS, and PhD.

To find the interesting rule based on above explanation then formula (1) will be used to measure concepts in generalization process against concept tree in order to find interesting rule and will be run on each of attribute in process selection generalization process as the sixth strategy step. The simple value as the most interesting value, the highest value as further generalization and the next bigger value as further simplification for unioning or grouping by attribute.

$$(\sum_{i=1}^{n} CR_i/CT_i)/n = (CR_1/CT_1 + ... + CR_n/CT_n)/n \quad (1)$$

where :

n = Maximum depth /level of concept tree

CR_i = Amount distinct concepts per level in attribute

CT_i = Amount concepts per level in Concept Tree

Table 7 shows the depth and length of concept trees which refer concept hierarchy in [6]

Table 7. Depth and length of concept tree

Depth /level	Length / Amount of Concepts				Total Concepts
	category	major	birthplace	GPA	
1	7	11	11	40	**69**
2	2	2	5	4	**13**
3	1	1	2	1	**5**
4			1		**1**

Total Concepts	10	14	19	45	88

For the next explaining will strengthen our approach with variance variable CR as amount concepts in generalization process. The same as before the generalization process for finding characteristic rule for graduate student with threshold value 2 but with different CR as amount of concepts.

Table 8 an example table which shows the result program as shown in figure 5. As a result the highest value as further generalization is birthplace attribute with formula value 0.682 and the unioning based on the next bigger value is major attribute with formula value 0.546 and the interesting attributes is GPA attribute with formula value 0.217 as the lowest value. Thus, the further generalization is on birthplace attribute and unioning on major attribute and table 3 as the interesting generalization relation.

Table 8. Amount of distinct concepts per level attribute for graduate characteristic with threshold=2

	Major			birthplace				GPA		
Depth/Level →	**1**	**2**	**3**	**1**	**2**	**3**	**4**	**1**	**2**	**3**
CR=Amount concepts	7	2		8	5	2		6	2	
CT=Amount concepts	11	2	1	11	5	2	1	40	4	1
CR/CT	0.636	1	0	0.727	1	1	0	0.15	0.5	0
Σ(CR/CT)/n	1.636/3=0.546			2.727/4=0.682				0.65/3=0.217		

For suppose there is the same highest formula value for attribute major and birthplace as shown in table 9, then there is a problem to decide attribute for further generalization and unioning because of equality formula value. Decision will be based on previous term where simple wide or length and depth or level of concept tree value will have a simple generalization process but in other hand many wide or length and depth or level of concept tree value will have many generalizations processes. As a result because birthplace attribute has 4 levels which more than major attribute with 3 levels then further generalization is on birthplace attribute with formula value 1 and the unioning on the next bigger value is on major attribute with formula value 1 and the interesting attributes is on GPA attribute with formula value 0.75, table 3 for example the result.

Table 9. Formula execution where there are the same highest formula value

	Major			birthplace				GPA		
Depth/Level →	**1**	**2**	**3**	**1**	**2**	**3**	**4**	**1**	**2**	**3**
CR=Amount concepts	11	2	1	11	5	2	1	10	4	1
CT=Amount concepts	11	2	1	11	5	2	1	40	4	1
CR/CT	1	1	1	1	1	1	1	0.25	1	1
Σ(CR/CT)/n	3/3=1			4/4=1				2.25/3=0.75		

If suppose the equality value happens on the same level attribute as shown in table 10 where major and GPA attribute have the same level then based on previous term where simple wide or length and depth or level of concept tree value will have a simple generalization process, then the selection will be decided based on multiplication non zero Amount distinct concepts (CR). The highest value multiplication concepts will act as further generalization and the next value as unioning and as result further generalization on GPA attribute where it has result 160 for multiplication 40*4*1, unioning on major attribute where it has result 22 for multiplication 11*2*1 and the interesting attribute is birthplace with formula value 0.859, table 5 for example the result.

Table 10. Formula execution where there are the same level and highest formula value

	Major			birthplace				GPA		
Depth/Level →	**1**	**2**	**3**	**1**	**2**	**3**	**4**	**1**	**2**	**3**
CR=Amount concepts	11	2	1	7	4	2	1	40	4	1
CT=Amount concepts	11	2	1	11	5	2	1	40	4	1
CR/CT	1	1	1	0.636	0.8	1	1	1	1	1
Σ(CR/CT)/n	3/3=1			3.44/4=0.859				3/3=1		

If suppose the equality has the same level and multiplication result as shown in table 11 where major and GPA attributes have the same level and multiplication result, then the selection will be decided based on the left or the first attribute. As a result further generalization on major attribute where it has result 22 for multiplication 11*2*1 as the first attribute, unioning on GPA attribute where it has result 22 for multiplication 2*11*1 as the last attribute and the interesting attribute is birthplace with formula value 0.859, table 2 for example the result.

Table 11. Formula execution where there are the same level and multiplication result

	Major			birthplace				GPA		
Depth/Level →	**1**	**2**	**3**	**1**	**2**	**3**	**4**	**1**	**2**	**3**
CR=Amount concepts	11	2	1	7	4	2	1	2	11	1
CT=Amount concepts	11	2	1	11	5	2	1	2	11	1
CR/CT	1	1	1	0.636	0.8	1	1	1	1	1
Σ(CR/CT)/n	3/3=1			3.44/4=0.859				3/3=1		

The previous equality value example is on the highest formula value and table 12 is an example when the equality is on the lowest formula value. Based on previous guidance then further generalization on birthplace attribute as highest formula value 0.75 and unioning on GPA attribute with formula value 0.667 which has the highest multiplication amount distinct concepts 160 for multiplication 40*4*1. The interesting attribute is major with formula value 0.667 which has the same value with GPA attribute but has less value multiplication amount distinct concept 22 for multiplication 11*2, table 4 for example the result.

Table 12. Formula execution where there are the same result at lowest value formula

	Major			**birthplace**				**GPA**		
Depth/Level →	**1**	**2**	**3**	**1**	**2**	**3**	**4**	**1**	**2**	**3**
CR=Amount concepts	11	2	0	11	5	2	0	40	4	0
CT=Amount concepts	11	2	1	11	5	2	1	40	4	1
CR/CT	1	1	0	1	1	1	0	1	1	0
Σ(CR/CT)/n	2/3=0.667			3/4=0.75				2/3=0.667		

4. REFERENCES

[1] Han, J., Cai, Y., Cercone, N. and Huang, Y. 1995. Discovery of Data Evolution Regularities in Large Databases. Journal of Computer and Software Engineering, 3(1), 41-69.

[2] Han,J., Cai, Y., and Cercone, N. 1993. Data-driven discovery of quantitative rules in relational databases. IEEE Trans on Knowl and Data Engin, 5(1), 29-40.

[3] Han,J. and Fu, Y. 1995. Exploration of the power of attribute-oriented induction in data mining. in U. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy, eds. Advances in Knowledge Discovery and Data Mining, 399-421, AAAI/MIT Press.

[4] Han, J., Cai, Y. and Cercone, N. 1992. Knowledge discovery in databases: An attribute-oriented approach. In Proceedings 18th International Conference Very Large Data Bases, Vancouver, British Columbia, 547-559.

[5] Han,J. 1994. Towards efficient induction mechanisms in database systems.Theoretical Computer Science, 133(2), 361-385.

[6] Cai, Y. 1989. Attribute-oriented induction in relational databases. Master thesis, Simon Fraser University.

[7] Fudger, D. and Hamilton, H.J. 1993. A Heuristic for Evaluating Databases for knowledge Discovery with DBLEARN. In Proceedings of the International Workshop on Rough Sets and Knowledge Discovery: Rough Sets, Fuzzy Sets and Knowledge Discovery, 44-51.

[8] Han, J. and Fu, Y. 1994. Dynamic Generation and Refinement of Concept Hierarchies for Knowledge Discovery in Databases. In Proceedings of AAAI Workshop on Knowledge Discovery in Databases, 157-168.

[9] Han, J., Fu, Y., Huang, Y., Cai, Y., and Cercone, N. 1994. DBLearn: a system prototype for knowledge discovery in relational databases, ACM SIGMOD Record, 23(2), 516.

[10] Han, J., Fu, Y., and Tang, S. 1995. Advances of the DBLearn system for knowledge discovery in large databases, in Proceedings of the 14th international Joint Conference on Artificial intelligence, 2049-2050.

[11] Han,J., Fu, Y.,Wang, W., Chiang, J., Gong, W., Koperski, K., Li,D., Lu, Y., Rajan,A., Stefanovic,N., Xia,B. and Zaiane,O.R. 1996. DBMiner:A system for mining knowledge in large relational databases. In Proceedings Int'l Conf. on Data Mining and Knowledge Discovery, 250-255.

[12] Han, J., Chiang, J. Y., Chee, S., Chen, J., Chen, Q., Cheng, S., Gong, W., Kamber, M., Koperski, K., Liu, G., Lu, Y., Stefanovic, N., Winstone, L., Xia, B. B., Zaiane, O. R., Zhang, S., and Zhu, H. 1997. DBMiner: a system for data mining in relational databases and data warehouses. In Proceedings of the 1997 Conference of the Centre For Advanced Studies on Collaborative Research, 8.

[13] Hilderman, R.J. and Hamilton, H.J. 2001. Knowledge Discovery and Measures of Interest. Kluwer Academic Publishers, Norwell, Massachusetts, USA.

MIDI Composition Tools Using JFugue Java API

Kartika Gunadi
Universitas Kristen Petra
Siwalankerto 121
Surabaya
(62) 031 2983000
kgunadi@petra.ac.id

Liliana
Universitas Kristen Petra
Siwalankerto 121
Surabaya
(62) 031 2983000
lilian@petra.ac.id

Hendra Kurnia Wijaya
Universitas Kristen Petra
Siwalankerto 121
Surabaya
(62) 031 2983000

ABSTRACT
In order to compose a song, composer needs to have many kinds of instruments to produce his composition. Therefore, an assisting tool that can represent instruments to compose a song is necessary. This software was developed to provide that assisting tool. This software was developed using JFugue Java API which could represent music in the form of programming language. The software used the Object Oriented Programming (OOP) concept and was programmed with java and NetBeans IDE 5.5 as the compiler. This software provides 16 tracks, 127 type of instruments, note setting (including octave, duration, and chord), and tempo setting. It is also can read, write, and save file in MIDI format.

Keywords
Composer, JFugue, Music.

1. INTRODUCTION
In designing and creating a song, a song composer needs various music instruments to produce his music composition. The limitation of the various music instruments availability will be an obstacle. In this case, we need a tool to simulate music instruments' voice. This tool is very useful in helping us to create a music composition.

In this research we try to produce an application program, MIDI Composition Tools which can be used to not only represent various music instruments, but also play the created music composition.

JFugue is an open source Java API (Application Programming Interface) without MIDI's complexity. JFugue represent music in programming language form [6]. JFugue supplies feature such as Music String to write musical note, harmonic musical note, music instruments, duration, and track, MIDI file operation.

Originally, JFugue represent music in string object which contain 11 music instructions as mention in Table 1.

Table 1. Music string

Music string	Explanation
Tone and brake	C, D, ….
Sharp, flat, Neutral	Sharp, flat, neutral
Octave	0-10
Chords	complete
Duration	1 – 1/128
Melody and harmony	
Tier (slur)	
Measure	Default: instrument piano
Instrument	128
Track/ channel	16
Tempo	0 – 255 default 120

2. DESIGN AND IMPLEMENTATION
This MIDI Composition Tools aplication program contains four sections as shown in the main aplication form in Figure 1.

- *Compose*, receive an input from user, translate it become a pattern in Music String form.
- *Create* MIDI *File*, to change Music String into MIDI *File*.
- *Load* MIDI *File*, to read MIDI file and change it into a pattern in Music String.
- *Play Music*, to play music, both pattern produced by *Composer* section and pattern produced by *Load* MIDI *File* section.

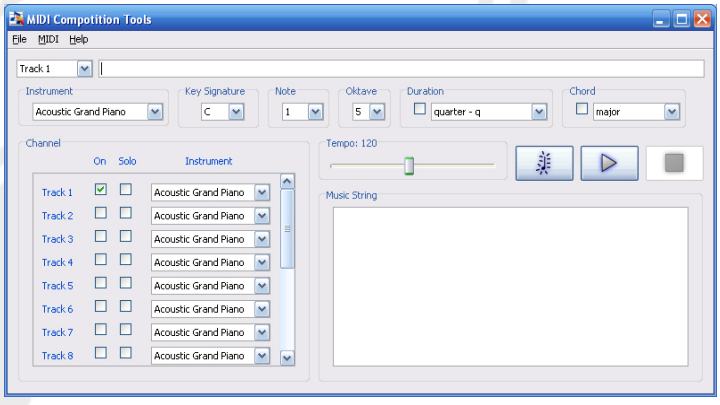


Figure 1. Main form

The main c*lasses* used in this application program are:

- *Pattern*, *class* which has a function to receive an input in string form. User pass this input and then the program will convert it become Music String.
- *Player*, *class* which has a function to convert Music String produced by class Pattern becomes audio signal. Beside function to convert Music String, this class also has a function to save and to read MIDI file. This class use library file from java, java.io.* and javax.sound.midi.*
- TrackOb, *class* which used to arrange *Music String*, *on/off track* state, and *instrument* from a *track.*
- NotOb, *class* which used to arrange *musical note* will be added in a *track.*
- DurationOb, *class* which use to arrange the duration of musical note will be added in a *track.*
- ChordOb, *class* which use to arrange a chord addition in a *track.*

3. EXPERIMENTS

To run this software, Java 5.0 or the newer version is required as platform to compile java programming. This application is package as MIDI Composition Tools.jar. We had done two kinds of experiments, Input system and Music String system experiment and MIDI file system experiment.

Input (*Composer*) system experiment and *Music String* system

Music note filling is done by press the ”tombol not balok” button. Then, a musical note will be added into the music string. This button has a tooltips which will inform us this button’s task, as shown in Figure 2.

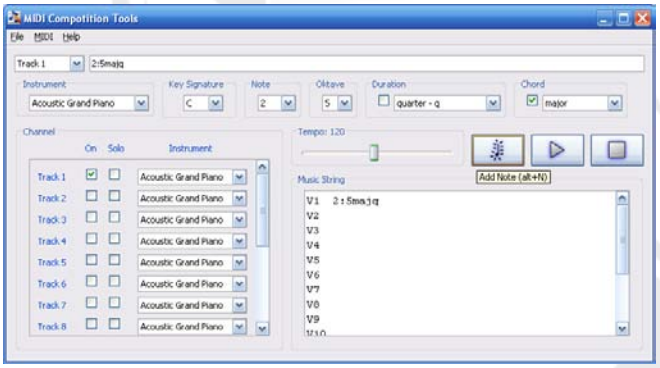


Figure 2. Input (composer) music string

MIDI File system experiment

This experiment had been done by saving the music string into MIDI file, reopening the MIDI file and playing it. This procedure can be done using the program or other player, as shown in Figure 3 and Figure 4.

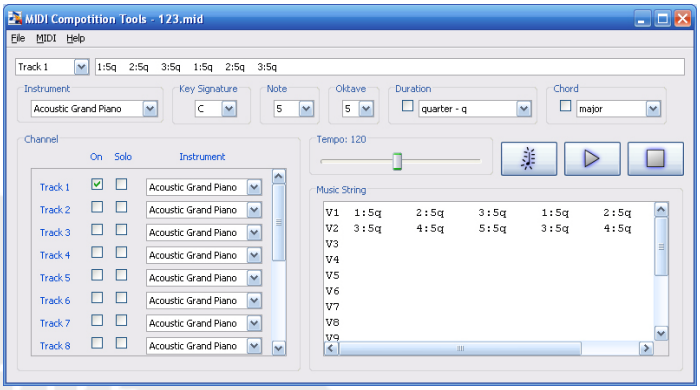


Figure 3. MIDI file reading

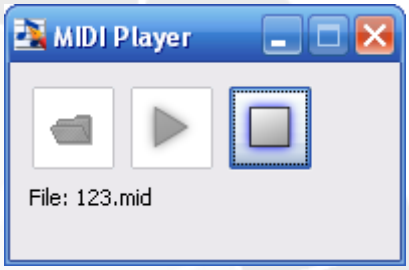


Figure 4. MIDI Player when *file* “123.mid” is sounded.

4. CONCLUSION

We had done some experiments and survey on the result. We conclude that JFUGUE JAVA API supports all features which needed in music composition, except double-sharps and double-flats. The usage of differ music instruments can be done easily by copying the music string. Then, we can edit the music string according to the chosen music instrument’s characteristic. After the editing process, it put in different channel. The tenth channel (V9) is the only channel which can produce non-chromatic percussion voice, drum.

5. REFERENCES

[1] Erckel, Bruce 2007. Thinking in Java, 2nd Edition. April 9, 2007. <http://www.campusrox.com/sourcecode/sc/TIJ2.pdf>

[2] Farrel, Joyce 1999. Java Programing : Introduction. Canada: Course Technology.

[3] Fikri, Rijalul. Adam, Ipam F. Prakoso, Imam 2004. Pemrograman Java. Yogyakarta: Penerbit ANDI

[4] Forster, Edward M. DIETEL 2007 - Java How to Program Fourth Edition. April 9, 2007. <http://www.dietel.com>

[5] Indrajani & Martin 2004. Pemrograman Berorientasi Objek dengan Java. Jakarta: PT Elex Media Komputindo

[6] Koelle, David 2007. JFugue Java API for Music Programming. April, 10, 2007. http://www.jfugue.org/howto.html

Mobile-based Interaction Using Dijkstra's Algorithm for Decision-making in Traffic Jam System

Puji Sularsih
Sarmag Program, Informatics Department
Gunadarma University
Jl. Margonda Raya 100 Pondok Cina, Depok 16424, West Java, Indonesia
eguy@student.gunadarma.ac.id

Egy Wisnu Moyo
Sarmag Program, Informatics Department
Gunadarma University
Jl. Margonda Raya 100 Pondok Cina, Depok 16424, West Java, Indonesia
dearest_v3chan@student.gunadarma.ac.id

Fitria H Siburian
Sarmag Program, Informatics Department
Gunadarma University
Jl. Margonda Raya 100 Pondok Cina, Depok 16424, West Java, Indonesia
ddxq_cuayang@student.gunadarma.ac.id

Sigit Widiyanto
Sarmag Program, Informatics Department
Gunadarma University
Jl. Margonda Raya 100 Pondok Cina, Depok 16424, West Java, Indonesia

Dewi Agushinta R
Sarmag Program, Informatics Department
Gunadarma University
Jl. Margonda Raya 100 Pondok Cina, Depok 16424, West Java, Indonesia
dewiar@staff.gunadarma.ac.id

ABSTRACT

Traffic jam detection system is a kind of system that can detect traffic in multiple locations. It requires the interaction with a system that requires an algorithm to meet the specific requirements. This paper presents a mobile-based interaction for detecting traffic jam using decision support system. The objectives of this application are to give the user the better online information of traffic flow of whole Jakarta through mobile application and to assist the user for making the decision of choosing an appropriate road. So, the user or the one who needs the information of the current situation on specific roads in Jakarta will be aimed by utilizing this application.

In this paper, we propose a mobile-based interaction for detecting traffic jam to provide decision-making using Dijkstra's shortest path algorithm. This algorithm will be used to establish the application in order to set the shortest path that can be passed by the traveler or the user. All of the paths will display the information of various circumstances in which can determine the present state of traffic. This information directly assists the user to avoid traffic congestion and to make a decision to choose the appropriate road.

Keywords

Decision Support System, Dijkstra's shortest path algorithm, Mobile-based Interaction, Traffic Jam

1. INTRODUCTION

An increasing number of vehicles and imprecise controls of traffic in Jakarta have become major issues that create congestion. The traffic congestion causes loss in productivity, consumes a lot of gasoline, diminishes air quality, creates a variety of safety hazards, often discourages tourism, and reduces business information [1]. All of these problems are required to be solved so that the traffic congestion can be reduced.

There are several different types of solution that have been taken to reduce traffic congestion in Jakarta, such as employs traffic policemen in important traffic points, attempt to lay more pavements to avoid congestion, etc. But with the advent of technology and increment of traffic flow, several approaches with less involvement of human have been taken. Contemporary approaches emphasize better information and control to use the existing infrastructure more efficiently [2]. In contemporary approaches, image processing, computer vision or robot vision, etc are highly recommended. In these types of solutions, involvement of computers provide many promising approaches because information feed through mobile applications or web networks can simply provided. Because of this, we are proposing an innovative method in detecting traffic congestion using mobile application.

In Indonesia, we can get the online information of traffic flow on certain location through some websites. One of the websites is http://lewatmana.com. In this website, the information of traffic flow will be obtained through cameras that are placed in important traffic points. This information will be updated every two hours. Therefore, the people who are connected to the internet network can use this application and choose an appropriate road to avoid congestion.

In this paper, we want to describe mobile-based interaction as a new generation of traffic jam detection system that has

tremendous potential to improve decision support system. This application will support real-time maps that can be viewed or accessed and provide the information of the current situation on specific roads in Jakarta. It also presents a better management decision making to the user or the traveler. So, the user not only can avoid traffic congestion but also can choose the appropriate road with a mobile phone. This new and improved application is necessary to develop an innovative traffic jam detection system.

This paper is structured as follows. The second section presents the path (graph theory) and the algorithm, Dijkstra's shortest path algorithm, that we used to establish the mobile application. The third section discusses the modeling of mobile-based interaction for detecting traffic congestion. The fourth section shows the final comments and conclusions.

2. RELATED WORKS

There are many approaches that can be used to solve the shortest path problem. Some researches have been using various algorithms, such as Dijkstra's shortest path algorithm. The Dijkstra's shortest path algorithm is still considered strong. The development of this algorithm is also still continued [5]. There are several approaches to perform the development of this algorithm through a hierarchical model [1]. So that, this algorithm can be adjusted to the real-weighted undirected graph. The Dijkstra's shortest path algorithm have an ability to create an efficient processing phase with a linear structure for single-source the shortest path through computation time O (m log a). Therefore, this algorithm has a powerful potential to solve the problems of traffic congestion which form a path which is the collection of graphs.

The traffic network can be supported by Arc-flag approach [2], which is one of result of the development of The Dijkstra's shortest path algorithm. This approach will be done by doing the partition through a graph. Then, each region will be marked. The mark describes that the region has a maximum number of edges of the graph. This kind of concept can be used to solve the traffic congestion problem in Jakarta through the unstructured road development. So, we can apply the main path in every certain location that has a number of edges.

For the certain locations that have a potential to cause traffic congestion can be presented using graph cut. The graph cut can be done through computing min-marginals approach. It will be implemented by labeling the calculated random field path efficiently based on dynamic graph [3]. This algorithm leads us to get a polynomial running time. However, planar algorithm [7] is also quite promising. The condition of highway in Jakarta overlapping must be changed into two-dimensional planar shape. After that we can do the cutting edge.

The algorithm that is used to search the shortest path also has to be described in a real map. To achieve this algorithm, we required a technique that can be performed with the Scalable Vector Graphics and Tiny Line SVG approaches for flexible display [8]. This approach supports generalization on the schematic map. By using the technique and combination software such as J2ME as software that supports connected limited device configuration [9], can be applied to the developed mapping application model.

3. METHODOLOGY

3.1 Data and Process

To build this application, we need information of traffic flow. The information can be obtained through cameras that are placed in important traffic points, such as Arteri Pondok Indah, Mampang, Jati Bening, etc. All of the places that have been mentioned have a potential to cause traffic congestion. The data of current situations on specific roads of whole Jakarta will be sent to the server to be processed and display through interface form or in the mobile application device. This interface will consists of specific parameters of congestion level in Jakarta. The illustration of the data process can be seen in figure 1.

Figure 1. Flow of the data process.

3.1.1 Redesign Map

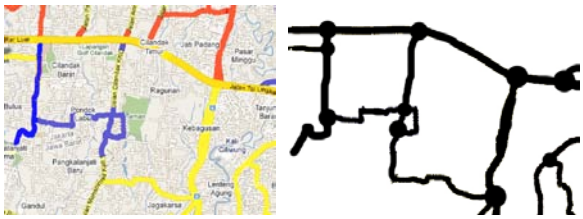
Google maps application is quite enough to describe the information of current situation on specific roads in Jakarta. However, the Google maps of the city of Jakarta have to be redesign so that it could compatible to our new application. If Google maps only give the information of certain roads, we have to redesign the information by changing the interface of the maps with using different colors as parameters of traffic congestion. We will use four different colors, such as red, blue, yellow, and purple. The color red is used to describe the state of traffic jam in which the vehicles are stopped. The color blue is used to show the certain roads that the vehicles are moving very slowly. The color yellow shows that there are an excessive number of vehicles on road. And the color purple that we used is to describe a condition of less number of vehicles on road. In this paper, the density range of vehicles on road will not be included as parameters of traffic congestion. We will only describe each edge with certain condition through its specific color .

3.1.2 Conversion

After redesigning the maps, we are going to change it into a graph, except in the interface, to represent a realistic pattern of certain roads in Jakarta. Then, the Dijkstra's shortest path

algorithm will be used to determine the shortest path from the graph. In our new application, as mentioned earlier, each graph will be synchronized to the real condition of roads in Jakarta. For instance, we will manually describe the crossroad as a node and the road as an edge that describe in figure 2.

The value of a single edge is based on the length of the path. As same as the graph theory, each of the edge has its own value. In this case, we give the biggest value to the longest path (road).



a) b)

Figure 2. Conversion the map to be a graph form. a) The Cilandak map from maps.google.com. b) The graph result from conversion map.

Each of the edge represents the certain road that has a possibility to be passed by the vehicles. Besides, we also need to convert the highway and the divided highway into a graph. It can be divided into the directed and undirected graph. We will convert the highway into directed graph or digraph. A directed graph or digraph is a pair G = (u, v). It means that an edge is related with two vertices (nodes) and considered to be directed from U to V. Meanwhile, the divided highway will be converted into undirected graph. The undirected graph is a graph in which edges have no orientation. It means that the edges are not ordered pairs, but sets {u, v} of vertices. The roads in Jakarta can be converted into undirected graph.

For the red path, as we mentioned earlier, the traffic jam is the unavoidable part of road. The red path will be removed from the path. We propose cut edges method. We choose the cut edges method in order to stabilize a number of different sub graphs. Furthermore, the cut edges method also can quest optimum solution for each function in polynomial time [4].

3.1.3 Quantification

The quantification process is used to represent the condition of the roads in whole city into quantitative value. Every condition that has various degree of the traffic jam will be quantified into different values. The quantitative value will be valuable according to the colors, such as the color purple is become the first value, the color yellow is become the second value, etc. For the color red, traffic jam condition, it can't be included into the graph. We provide the quantification process using cut edges algorithm [11].

3.2 Algorithm Development

To apply the shortest path, we use one of greedy method, Dijkstra's shortest path algorithm, where has been required to input value of the first node as initial position and value of the second node as destination. Normally, the shortest path can be obtained from the shortest distance of road and the current situation of road without traffic congestion. We multiply the distance of each edge with the quantification result. Thus, the value of the path can be provided using our formula.

$$W_i = E_i \times C_n \quad (1)$$

$$P_i = \sum_{i \in K} W_i \quad (2)$$

Formula (1) shows W, the weight of each selected edge, which is depended by i. The value of i is the variable array that defines the index of each edge. W is calculated by multiplying E as a length of the edge between the cross road by C_n, the constant number that is depended by n as the index of constant number. The index of constant number will be defined as the condition which have been quantified, from 1 until 4. The condition is retrieved from the quantification process. The purple color has value 1, yellow color has value 2, blue color has value 3, and red color has value 4. So, the value of C_n depends on the condition of the way.

Formula (2) is used to count the value of sub path that will be used to find the shortest path. P_i is the value of each path that will be selected. The sigma symbol means the iteration of W_i that is determined by K. The K value is the total number of selected vertexes. It is the subset of all of the vertexes in the graph.

In this paper, we modify the Dijkstra's shortest path algorithm to obtain the three selected shortest paths in order to provide the decision making and risk evaluation to the users. We assume that the users will not obviously take the first selected shortest path. The users probably will consider to take another path to avoid the traffic congestion. Therefore, we provide two other paths that can be chosen. When one of the path has been selected, the program will be repeated. However, the program will remove one of the edge that is connected to the node which has other alternative edges in the selected path. The program will repeat until the initial node can not be passed through its edge. The resulting value of each path can be obtained through formula (2).

After we get the value of each path, we will specify the three minimum values in order to provide the best three paths that can be passed. The best three paths will be processed to obtain the percentage of each selected path through its edges. We will generate the paths that have the same condition, such as, the paths which have value 2. Then, we can determine the percentage of each constant number by multiplying the length of the edge by the constant number and then compare the result directly with the value of the each selected path.

$$P_1 = W_1 + W_2 + W_3 + \cdots + W_K \quad (3)$$

$$P_2 = W_1 + W_2 + W_3 + \cdots + W_K \quad (4)$$

$$P_3 = W_1 + W_2 + W_3 + \cdots + W_K \quad (5)$$

$$P_1 = E_1 \times C_1 + E_2 \times C_1 + E_3 \times C_2 + \cdots + E_i \times C_n \quad (6)$$

$$RC_m = \frac{\sum_{i=1} E_i \times C_m}{P_t} \times 100\% \quad (7)$$

In words, in order to obtain RC, the percentage of each constant number, we will multiply the total number of E_i, the length of the edge, by C_m, the constant number, and then devide the result with the value of the each selected path. *m* represents the index of the condition of each selected path. *m* is the subset of *n*. *L* is the total number of selected edges which have the same constant number.

We can provide three paths (shortest paths) to the user from the best of three combinations that have been chosen, but not always there are three suggestions, if the algorithm only find 2 or less the suggest path, the visual only display 1 or 2 way. It depends on the condition of traffic. The following pseudo-code gives a brief description of the working of the Dijkstra's shortest path algorithm [10].

Procedure Dijsktra (V: set of vertices 1... n *{Vertex 1 is the source}*
Adj[1…n] of adjacency lists;
EdgeCost(u, w): edge – cost functions;)
Var: sDist[1…n] of path costs from source (vertex 1);
{sDist[j] will be equal to the length of the shortest path to j}
Begin:
Initialize
{Create a virtual set Frontier to store i where sDist[i] is already fully solved}
Create empty Priority Queue New Frontier;
sDist[1]←0; *{The distance to the source is zero}*
forall vertices w in V – {1} **do** *{no edges have been explored yet}*
sDist[w]←∞
end for;
Fill New Frontier with vertices w in V organized by priorities sDist[w];
endInitialize;

repeat
v←DeleteMin{New Frontier}; *{v is the new closest; sDist[v] is already correct}*
forall of the neighbors w in Adj[v] **do**
if sDist[w]>sDist[v] +EdgeCost(v,w) then
sDist[w]←sDist[v] +EdgeCost(v,w)
update w in New Frontier *{with new priority sDist[w]}*
endif
endfor
until New Frontier is empty
endDijkstra;

There are many crossroads in Jakarta. It means that the program will process many nodes to gain the shortest path. So that, the computational complexity tends to be very difficult to be defined. To increase the computational process, we will restrict the region of the graph through the Arc-Flag Approach [5].

3.3 Visualization

We plan to build our application using Java 2 Mobile Edition (J2ME). The combination of *Connected Limited Device Configuration* (CLDC) and *Mobile Information Device Profile* (*MIDP)* can provide a solid Java platform for developing applications to run on devices with limited memory, processing power, and graphical capabilities. So, the application that developed can run in the small network like GPRS or 3G. CLDC defines the base set of application programming interfaces devices like mobile phones, pagers, and mainstream personal digital assistants. We can use these privileges to optimize the design and application systems.

The design, which is the first form of our application, can be shown in figure 3. In this part, we display a map that can be a random map, with image magnification capability. The capability of obtaining magnifications is 1 to 3 times. We only offer to magnify the image by three times because the mobile applications have a limited ability in memory and loading data through the GPRS or 3G network.

a)

b)

c) d)

Figure 3. Design our application. a) The first preview with index parameter. b) The first shortest path. c) The second shortest path. d) The third shortest path.

The right side in our application will show the information of index or parameters of traffic congestion through various colors, as we mentioned earlier, the color red, blue, yellow, and purple. While the lower part will give a search facility, it is used to search an alternative path (road). The user can enter the initial position, the place where the user is located, and the final destination. Every data that has been input will be sent to our server through internet network. Then, all of the data, the certain location and its current situation (through the camera and other data from internet network) will be processed using Dijkstra's shortest path algorithm.

In the second form of our application, we propose the result which is the three best paths (roads) with the optimum way that can be passed by the user. Three best paths will be showed with green color. The first path is a recommended road that the system suggested to the user. It is the shortest path without congestion or the highest value. Then, the following path, the system will suggest another path that has lower value through the algorithm. In each selected path, we will describe the percentage of path (road) condition which is including the color blue, yellow, and purple. For example, the color blue has 0%, the color yellow has 100%, and the color purple has 0%. It means that this path shows the condition of an excessive number of vehicles on road. . Every path in this system can be visualized using the map application by choosing one of the paths (click the navigation button in the center).

4. CONCLUSIONS

The Dijkstra's shortest path algorithm is very useful in resolving problems of traffic congestion. Combining the algorithm with map visualization techniques and software that supports the CLDC can provide visualization effects and flexible system.

Our new application of detecting traffic jam can be accessed and implemented in mobile application through internet network. It can assist to provide the information about traffic points with traffic jam and facilitate the users to choose the appropriate alternative path (route) to avoid congestion. Using GPRS, this application can help the users to save their valuable time and reduce the cost for using the GPS.

Further development of our planning application is the development of a database system and the connection with the server devices to the mobile application interface. The implementation of our application will be done in real-time data according to the initial design.

5. REFERENCES

[1] El-Geneidy, A.M, Ayad, H.M., El-Baghdady, N.S., and Azzam, Y.A. A Decision Support System For Land Use Activity Changes Using Data Gained from The Intelligent Transportation Systems.

[2] Beymer, D., McLauchlan, P., and Malik, J. 1998, 1 December. A Real-Time Computer Vision System for Vehicle Tracking and Traffic Surveillance. Transportation Research-C.

[3] Chen, M., Chowdhury, R.A., Ramachandran, V., Roche, D.L., and Tong, L. 2007. Priority Queues and Dijkstra's Algorithm. UTCS Technical Report TR-07-54 (October 12, 2007).

[4] Pettie, S. and Ramachandran, V. 2005. A Shortest Path Algorithm For Real-Weighted Undirected Graphs. Society for Industrial and Applied Mathematics. SIAM J. COMPUTE. C (Vol. 34, No. 6, pp. 1398–1431).

[5] M"ohring, R.H., Schilling, H., Sch¨ utz, B., Wagner, D., Willhalm, T. 2005. Partitioning Graphs to Speed Up Dijkstra's Algorithm. 4th International Workshop on Efficient and Experimental Algorithms.

[6] McMahan, H.B. and Gordon, G.J. 2005. Generalizing Dijkstra's Algorithm and Gaussian Elimination for Solving MDPs. DARPA's MICA project, AFRL contract F30602-01-C-0219.

[7] Boyer, J.M. and Myrvold, W.J. 2004. On the Cutting Edge: Simplified O(n) Planarity by Edge Addition. Journal of Graph Algorithms and Applications. http://jgaa.info/ vol. 8, no. 3. pp. 241–273.

[8] Harun, H., Jailani, N., Yatim, N.F.M., Abu Bakar, M., Zakaria, M.S., and Abdullah, S. 2008. Kajian Terhadap Teknik Visualisasi Peta Aplikasi LBS Pada Peranti Mudah Alih. Fakulti Teknologi dan Sains Maklumat, Universiti Kebangsaan Malaysia.

[9] Oracle. Chapter 8: MIDP Aplication Model.

[10] Puthuparampil, M. Report Dijkstra's Algorithm.

[11] Kohli, P. and Torr, P.H.S. 2008. Measuring uncertainty in graph cut solutions. Computer Vision and Image Understanding 112. pp. 30–38.

Model and Boarding Simulation for Reducing Seat and Aisle Interferences Between Passenger

Bilqis Amaliah
Informatics Department,
Faculty of Information Technology,
Institut Teknologi Sepuluh Nopember
bilqis@if.its.ac.id

Victor Hariadi
Informatics Department,
Faculty of Information Technology,
Institut Teknologi Sepuluh Nopember
victor@its-sby.edu

Antonius Malem Barus
Informatics Department,
Faculty of Information Technology,
Institut Teknologi Sepuluh Nopember
Antonius.mb@gmail.com

ABSTRACT
The airline gets revenue while their aircraft are flying. There are many things that influence the plane on the ground, for example: the time needed for passengers to get off the plane, baggage loading and unloading, fueling, boarding time, etc. This paper presents a few strategic model boarding for reducing seat and aisle interference and for reducing boarding time. Mixed Integer Non Linier Programming is used for generating boarding model. ProModel is used for simulating and the result are time and sum of seat and aisle interferences. Airbus-320 is used to apply this simulation model. Some of the things that affect the boarding strategy model are the number of rows, number of groups and number of passengers that included for each group. The simulation result show that 6 group boarding model reducing 57,1% for interferences and 6,82% for boarding time over traditional pure back to front boarding model.

Keywords
Boarding, MINLP, transportation.

1. INTRODUCTION
For commercial airlines, one of the factors that determine the efficiency operational of aircraft is the transition time (turnaround time). the transition begins from arrival to departure of an airplane. Factors that influence the transition time on the aircraft included the time passengers to get off the aircraft, loading and unloading baggage, fueling, aircraft maintenance, boarding time etc.

Boarding time is one of the factor that can influence the efficiency operational of a flight. Boarding time is difficult to control by the flight service providers due to limitations in control of the passengers.

Because of that, it is necessary for the researcher to find out the optimal boarding strategies to improve the efficiency. After finding the optimal strategy then it is necessary to do simulations for modeling the boarding situation.

Research on boarding, has been carried out by several previous researchers. According to Van Landeghem and Beuselinck [3], many factors that determine the turnaround time, such as: loading and unloading of goods and passengers, checking passengers, fuel filling. In this paper also discussed some kind of boarding strategies. Bazargam [1] also discusses a few boarding model with linear programming approach.

This paper presents a few strategic model boarding for reducing seat and aisle interference and for reducing boarding time. Mixed Integer Non Linier Programming is used for generating boarding model. ProModel is used for simulating and the result are time and sum of seat and aisle interferences. Airbus-320 is used to apply this simulation model. Some of the things that affect the boarding strategy are the number of rows, number of groups and number of passengers that included for each group.

The purpose of this research is choosing the optimal boarding model, where the number of seat and aisle interference is minimum and so is the time.

Section 2, 3 and 4 define mixed Integer NonLinier Programming (MINLP), Neos Server and Airbus A-320. Section 5, 6 and 7 define seat interference, aisle interference and Penalty value. Section 8 and 9 formulated model boarding and design system. Section 10 experiment and implementation then finally section 11 concludes this paper.

2. MIXED INTEGER NONLINIER PROGRAMMING (MINLP)
Mixed Integer Nonlinear Programming (MINLP) is a variety of forms of Nonlinear Programming problems that combined with Integer Programming. MINLP is a natural approach to formulate the optimization problem [3].

Algorithm that can be used to solve the problem on mixed integer nonlinear programming is branch and bound algorithm [2].

3. NEOS SERVER
Solution for optimization problems with many variables (> 300) cannot be solved using AMPL student version. One solution to solve the problems with this many variables is to use Neos Server. Neos Server is a server that serves the optimization problem in a way to upload a file that contains the mathematical AMPL model optimization.

AMPL will read the model from *. mod files and data from the *. dat files and will be completed in accordance with the solver who has previously selected.

The file contains a model of mathematical models created by the programming language AMPL and Gams. Data files contain data that will be input to the model. As for modifying the output of the proposed solution, users can use command files [4]. These files can be uploaded via the website server Neos (http://neos.mcs.anl.gov/neos/).

Figure 1. *Website* NEOS server

MINLP Solver is a solver that available on the Neos Server, it solve mixed integer optimization problems with constrain (Mixed Integer Nonlinearly Constrained Optimization).

4. AIRBUS A-320

Airbus A320 is an aircraft that can accommodate 150 passengers, consisting of 12 passengers in business class and 138 passenger in economy class.

In general, the column marked letter A, B, C, D, E and F (economy class) and A, C, D and F (business class). For economy class, A and F is a seat near the window (window), B and E are the middle seats (middle), while C and D is an aisle seat (the aisle).

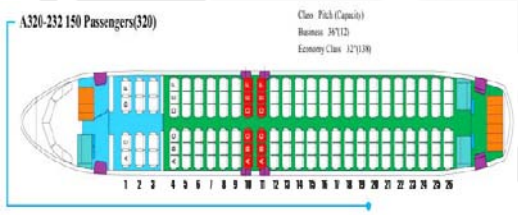


Figure 2. Layout kabin A320

Line in the cabin starts from 1 to 26, which consists of numbers 1 to 3 is the business class and 4 to 26 is the economy class.

Conventionally, boarding process is dividing passenger into groups. Boarding process will filled the seats from back to front [8]. In this study, this model is called the BF model (Back to Front). Where in this model, fill the back seat first can reduce the interference.

Variable N represents the set of rows and M = (A, B, C, D, E, F) represents the set of column. Given a number to each row i in N and j in M seat position, then each individual position chairs can be identified by using a pair (i, j).

By including the group's position on the chair, it can be established boarding strategy. For example the problems boarding the plane, if each pair (i, j) is inserted at the boarding group k, k in G which represents a set of groups. Further defining the decision variables $x_{i,\,j}$, k = 1 if the seat (i, j) be included in the group k and $x_{i,\,j,\,k}$ = 0 for values other than, where I in N, j in M and k in G.

5. SEAT INTERFERENCE

1. Seat Interferences

Seat interference is interferences that occur when passengers who will sit near windows and the passengers in middle or aisle already sit. [8]. If x indicates passengers who join a group, then the scenario that may occur are:

a. Three passengers were in the same group and will occupy a seat on the right or left (xxx). The model is as follows:

$$\sum_{i\in N}\sum_{k\in G} x_{i,A,k}x_{i,B,k}x_{i,C,k} + \sum_{i\in N}\sum_{k\in G} x_{i,D,k}x_{i,E,k}x_{i,F,k}$$

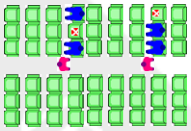


Figure 3. Seat (xxx)

b. Two passengers were in a group and followed by the other passengers in the after group (xx_x). If k, l in G where k <l, then the model is as follows:

$$\sum_{i\in N}\sum_{k,l\in G:k<l} x_{i,A,k}x_{i,B,k}x_{i,C,l} + \sum_{i\in N}\sum_{k,l\in G:k<l} x_{i,A,k}x_{i,B,l}x_{i,C,k}$$

$$\sum_{i\in N}\sum_{k,l\in G:k<l} x_{i,A,l}x_{i,B,k}x_{i,C,k} + \sum_{i\in N}\sum_{k,l\in G:k<l} x_{i,F,k}x_{i,E,k}x_{i,D,l} +$$

$$\sum_{i\in N}\sum_{k,l\in G:k<l} x_{i,F,k}x_{i,E,l}x_{i,D,k} + \sum_{i\in N}\sum_{k,l\in G:k<l} x_{i,F,l}x_{i,E,k}x_{i,D,k}$$

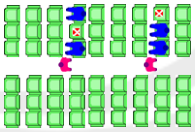


Figure 1. Seat (xx_x)

c. One passenger in a group for one section and row, followed by two passengers on the same line and in the later group (x_xx). The model is as follows:

$$\sum_{i\in N}\sum_{k,l\in G:k<l} x_{i,A,k}x_{i,B,l}x_{i,C,l} + \sum_{i\in N}\sum_{k,l\in G:k<l} x_{i,A,l}x_{i,B,k}x_{i,C,l} +$$

$$\sum_{i\in N}\sum_{k,l\in G:k<l} x_{i,A,l}x_{i,B,l}x_{i,C,k} + \sum_{i\in N}\sum_{k,l\in G:k<l} x_{i,F,l}x_{i,E,l}x_{i,D,k} +$$

$$\sum_{i\in N}\sum_{k,l\in G:k<l} x_{i,F,l}x_{i,E,k}x_{i,D,l} + \sum_{i\in N}\sum_{k,l\in G:k<l} x_{i,F,k}x_{i,E,l}x_{i,D,l}$$

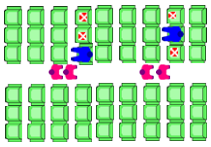


Figure 2. Seat (x_xx)

d. Three passengers in the group different are assigned a seat in the row and the same section (x_x_x). If k, l, m in G where k <l <m, then the model is as follows::

$$\sum_{i\in N}\sum_{k,l,m\in G\cdot k<l<m} x_{i,A,l}x_{i,B,m}x_{i,C,k} + \sum_{i\in N}\sum_{k,l,m\in G\cdot k<l<m} x_{i,A,l}x_{i,B,k}x_{i,C,m} +$$
$$\sum_{i\in N}\sum_{k,l,m\in G\cdot k<l<m} x_{i,A,m}x_{i,B,l}x_{i,C,k} + \sum_{i\in N}\sum_{k,l,m\in G\cdot k<l<m} x_{i,A,k}x_{i,B,m}x_{i,C,l} +$$
$$\sum_{i\in N}\sum_{k,l,m\in G\cdot k<l<m} x_{i,A,m}x_{i,B,k}x_{i,C,l} + \sum_{i\in N}\sum_{k,l,m\in G\cdot k<l<m} x_{i,F,m}x_{i,E,k}x_{i,D,l} +$$
$$\sum_{i\in N}\sum_{k,l,m\in G\cdot k<l<m} x_{i,F,k}x_{i,E,m}x_{i,D,l} + \sum_{i\in N}\sum_{k,l,m\in G\cdot k<l<m} x_{i,F,m}x_{i,E,l}x_{i,D,k} +$$
$$\sum_{i\in N}\sum_{k,l,m\in G\cdot k<l<m} x_{i,F,l}x_{i,E,k}x_{i,D,m} + \sum_{i\in N}\sum_{k,l,m\in G\cdot k<l<m} x_{i,F,l}x_{i,E,m}x_{i,D,k}$$

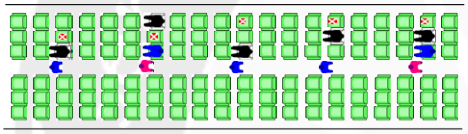


Figure 6.Seat (x_x_x)

6. AISLE INTERFERENCE

Aisle interference is interference that occurred in the aisle, while passengers who will sit on the next line was blocked by passengers who will sit on this line, because passengers who will sit on this line, put his baggage into kabin or seat interference occurs[8].

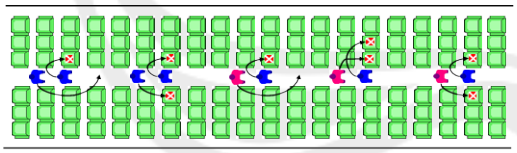


Figure 1. Aisle Interferences

If x and y are passengers on the different groups, Aisle Interference is as follows :

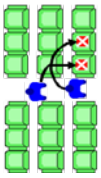
a. Passenger within group

Aisle Interference will happen, if minimal one passenger is blocked by other passenger in the same group. These scenarios are divided become three parts:

i. Passenger will sit at the same row and the same side (xy_sr_ss).

At this scenario, if u,v ∈ L, R where u ≠ v, then the model is as follows :

$$\sum_{k\in G}\sum_{i\in N}\sum_{u,v\in Lu\neq v}\left(x_{i,u,k}x_{i,v,k}\right) + \sum_{k\in G}\sum_{i\in N}\sum_{u,v\in Ru\neq v}\left(x_{i,u,k}x_{i,v,k}\right)$$



ii. Passenger will sit at the same row and the different site (xy_sr_ds).

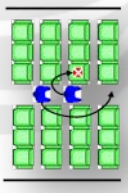
At this scenario, if u, v ∈ M where u ∈ L and v ∈ R, then the model is as follows:

$$2\sum_{k\in G}\sum_{i\in N}\sum_{u,v\in M:u\in L,v\in R}\left(x_{i,u,k}x_{i,v,k}\right)$$

iii. Passenger will sit at the next row (xy_hr).

This Skenario will happen when the next passenger in the same group will sit at the further sit. If a, b ∈ N, then the model is as follows:

$$\sum_{k\in G}\sum_{a,b\in N}\sum_{u,v\in M}\left(x_{a,u,k}x_{b,v,k}\right)$$



b. Passenger berween group

Aisle Interference will happen, if minimal one passenger is blocked by other passenger in the different group. These scenarios are divided become three parts:

i. Passenger will sit at the same row and the same side (xy_sr_ss).

At this scenario, if k, l ∈ G, where k < l, then the model is as follows:

$$\sum_{k,l \in G:k<l} \sum_{i \in N} \sum_{u,v \in R} (x_{i,u,k} x_{i,v,l}) + \sum_{k \in G:k<l} \sum_{i \in N} \sum_{u,v \in L} (x_{i,u,k} x_{i,v,l})$$

ii. Passenger will sit at the same row and the different site (xy_sr_ds). the model is as follows :

$$\sum_{k,l \in G:k<l} \sum_{i \in N} \sum_{u \in L, u \in R} (x_{i,u,k} x_{i,v,l}) + \sum_{k \in G:k<l} \sum_{i \in N} \sum_{u \in L, u \in R} (x_{i,u,l} x_{i,v,k})$$

iii. Passenger will sit at the next row (xy_hr). the model is as follows :

$$\sum_{k,l \in G:k<l} \sum_{a,b \in N:a<b} \sum_{u,v \in M} (x_{a,u,k} x_{b,v,l})$$

7. PENALTY VALUE

Penalty value is used to give weight to the value of mathematical models that will be made. In this case, to determine the penalty values is done by calculating the value of seat and aisle interference probability.

Table 1. Seat Penalty

Penalti	Susunan *Boarding*	E (*No. of interference*)
λ_1^s	[window, middle, aisle]	1,5
λ_2^s	[window, middle] → [aisle]	0,5
λ_3^s	[window, aisle] → [middle]	1,5
λ_4^s	[middle, aisle] → [window]	2,5
λ_5^s	[window] → [middle, aisle]	0,5
λ_6^s	[middle] → [window, aisle]	1,5
λ_7^s	[aisle] → [window, middle]	2,5
λ_8^s	[window] → [aisle] → [middle]	1
λ_9^s	[middle] → [window] → [aisle]	1
λ_{10}^s	[middle] → [aisle] → [window]	2
λ_{11}^s	[aisle] → [window] → [middle]	2
λ_{12}^s	[aisle] → [middle] → [window]	3

Table 2. Aisle Penalty

Penalti	Keterangan	E (No. of interference)
$\lambda_1^a, \lambda_2^a, \lambda_3^a$	Within group	$1/s_1$
$\lambda_4^a, \lambda_5^a, \lambda_6^a$	Between group	$1/(s_1\ s_2)$

8. MODEL BOARDING INTERFERENCE

Minimize

Z =

$$\lambda_1^s \sum_{i \in N} \sum_{k \in G} x_{i,A,k} x_{i,B,k} x_{i,C,k} + \lambda_1^s \sum_{i \in N} \sum_{k \in G} x_{i,D,k} x_{i,E,k} x_{i,F,k} +$$

... (xxx)

$$\lambda_2^s \sum_{i \in N} \sum_{k,l \in G:k<l} x_{i,A,k} x_{i,B,k} x_{i,C,l} + \lambda_3^s \sum_{i \in N} \sum_{k,l \in G:k<l} x_{i,A,k} x_{i,B,l} x_{i,C,k}$$

$$\lambda_4^s \sum_{i \in N} \sum_{k,l \in G:k<l} x_{i,A,l} x_{i,B,k} x_{i,C,k} + \lambda_2^s \sum_{i \in N} \sum_{k,l \in G:k<l} x_{i,F,k} x_{i,E,k} x_{i,D,l}$$

$$\lambda_3^s \sum_{i \in N} \sum_{k,l \in G:k<l} x_{i,F,k} x_{i,E,l} x_{i,D,k} + \lambda_4^s \sum_{i \in N} \sum_{k,l \in G:k<l} x_{i,F,l} x_{i,E,k} x_{i,D,k} +$$

... (xx_x)

$$\lambda_5^s \sum_{i \in N} \sum_{k,l \in G:k<l} x_{i,A,k} x_{i,B,l} x_{i,C,l} + \lambda_6^s \sum_{i \in N} \sum_{k,l \in G:k<l} x_{i,A,l} x_{i,B,k} x_{i,C,l} +$$

$$\lambda_7^s \sum_{i \in N} \sum_{k,l \in G:k<l} x_{i,A,l} x_{i,B,l} x_{i,C,k} + \lambda_7^s \sum_{i \in N} \sum_{k,l \in G:k<l} x_{i,F,l} x_{i,E,l} x_{i,D,k}$$

$$\lambda_6^s \sum_{i \in N} \sum_{k,l \in G:k<l} x_{i,F,l} x_{i,E,k} x_{i,D,l} + \lambda_5^s \sum_{i \in N} \sum_{k,l \in G:k<l} x_{i,F,k} x_{i,E,l} x_{i,D,l}$$

... (x_xx)

$$\lambda_{11}^s \sum_{i \in N} \sum_{k,l,m \in G:k<l<m} x_{i,A,l} x_{i,B,m} x_{i,C,k} + \lambda_9^s \sum_{i \in N} \sum_{k,l,m \in G:k<l<m} x_{i,A,l} x_{i,B,k} x_{i,C,m} +$$

$$\lambda_{12}^s \sum_{i \in N} \sum_{k,l,m \in G:k<l<m} x_{i,A,m} x_{i,B,l} x_{i,C,k} + \lambda_8^s \sum_{i \in N} \sum_{k,l,m \in G:k<l<m} x_{i,A,k} x_{i,B,m} x_{i,C,l} +$$

$$\lambda_{10}^s \sum_{i \in N} \sum_{k,l,m \in G:k<l<m} x_{i,A,m} x_{i,B,k} x_{i,C,l} + \lambda_{10}^s \sum_{i \in N} \sum_{k,l,m \in G:k<l<m} x_{i,F,m} x_{i,E,k} x_{i,D,l} +$$

$$\lambda_8^s \sum_{i \in N} \sum_{k,l,m \in G:k<l<m} x_{i,F,k} x_{i,E,m} x_{i,D,l} + \lambda_9^s \sum_{i \in N} \sum_{k,l,m \in G:k<l<m} x_{i,F,l} x_{i,E,k} x_{i,D,m} +$$

$$\lambda_{11}^s \sum_{i \in N} \sum_{k,l,m \in G:k<l<m} x_{i,F,l} x_{i,E,m} x_{i,D,k} +$$

... (x_x_x)

$$\lambda_1^a \sum_{k \in G} \sum_{i \in N} \sum_{u,v \in L:u \neq v} (x_{i,u,k} x_{i,v,k}) + \lambda_1^a \sum_{k \in G} \sum_{i \in N} \sum_{u,v \in R:u \neq v} (x_{i,u,k} x_{i,v,k})$$

$$2\lambda_2^a \sum_{k\in G}\sum_{i\in N}\sum_{u,v\in M\mathrm{u}\in L,v\in R}\left(x_{i,u,k}x_{i,v,k}\right)+\lambda_3^a \sum_{k\in G}\sum_{a,b\in N}\sum_{u,v\in M}\left(x_{a,u,k}x_{b,v,k}\right)+$$

… (within group)

$$\lambda_4^a \sum_{k,l\in G\text{-}k<l}\sum_{i\in N}\sum_{u,v\in R}\left(x_{i,u,k}x_{i,v,l}\right)+\lambda_4^a \sum_{k\in G}\sum_{i\in N}\sum_{u,v\in L}\left(x_{i,u,k}x_{i,v,l}\right)+$$

$$\lambda_5^a \sum_{k,l\in G\text{-}k<l}\sum_{i\in N}\sum_{u\in L,u\in R}\left(x_{i,u,k}x_{i,v,l}\right)+\lambda_5^a \sum_{k\in G}\sum_{i\in N}\sum_{u\in L,u\in R}\left(x_{i,u,k}x_{i,v,k}\right)+$$

$$\lambda_6^a \sum_{k,l\in G\text{-}k<l}\sum_{a,b\in N\text{-}a<b}\sum_{u,v\in M}\left(x_{a,u,k}x_{b,v,l}\right)$$

… (between group)

Subject to:

$$\sum_{k\in G} x_{i,j,k}=1:\ i\in N, j\in M$$

$$\sum_{k\in G}\sum_{j\in M} x_{i,j,k}\geq C_{min}:\ k\in G$$

$$\sum_{k\in G}\sum_{j\in M} x_{i,j,k}\leq C_{max}:k\in G$$

$$x_{i,j,k}\in\{0,1\}:i\in N, j\in M, k\in G$$

9. DESIGN SYSTEM

In this research, Neo Server is used to solve the optimization problem. Neo Server Results is a boarding model. This boarding model is analysis program input. Analysis program result the number of interferences and also binary model of boarding. Boarding binary model stored in the form of *. xls. *. xls file is input for the Pro Model simulation program. Pro model simulated 100 times and result average number of interferences and the average boarding time.

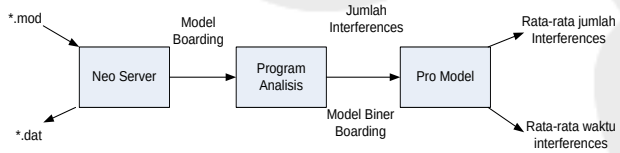


Figure 8. Design system

10. EXPERIMENT AND IMPLEMENTATION

The experiment is done by using analysis and simulation programs. The data is used in this experiment process is data boarding strategy model that has been generated from NEOS SERVER.

From the results of analysis and simulation shows that the solutions given from MINLP 6 model better than the BF 6 model (Back to Front).

From the calculation analysis, MINLP 6 reduces 43.89% number of interferences compare with BF 6 Model.

From the ProModel simulation shows that MINLP 6 reduce the number of interferences 57,1% better than BF 6 Model. And MINLP 6 Model reducing 6.82% boarding time better than BF 6 model.

From evaluations that have been presented, MINLP 6 model can be recommended as one of the alternative strategies to improve the efficiency of boarding time Airbus A320 aircraft. Futhermore, this model can be implemented in actual boarding system for Airbus A-320.

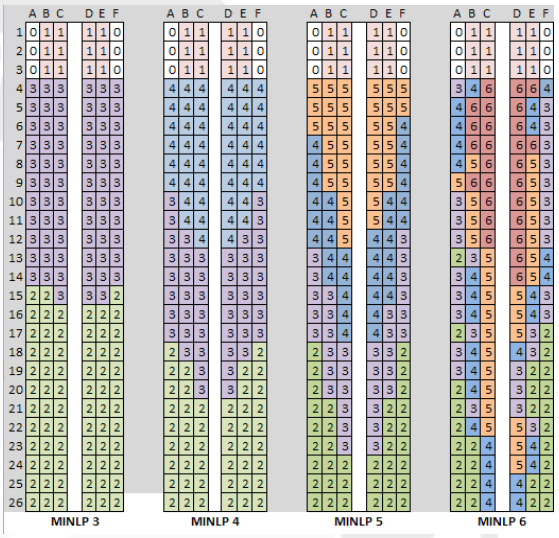


Figure 9. MINLP model

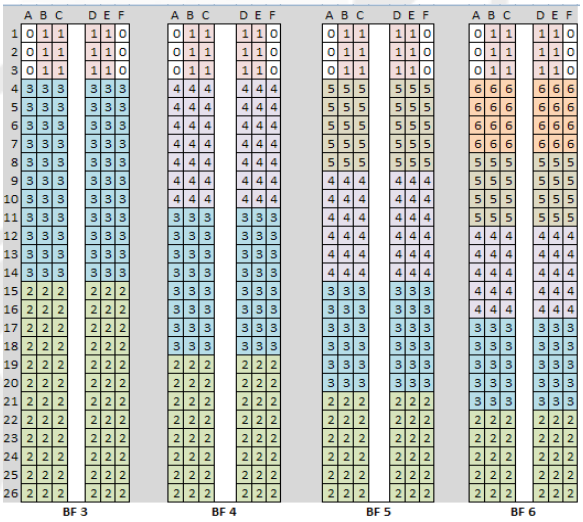


Figure 10. BF model (*Back to Front*)

Table 3. Result of model MINLP and BF using analisys program

		BF3	BF4	BF5	BF6	MIN LP3	MIN LP4	MIN LP5	MIN LP6
	Nr. Seat Interferences								
	Business	3	3	3	3	3	3	3	3

	(1xx)								
	Economy (xxx)	69	69	69	69	66	51	16,5	0
	xx_x	0	0	0	0	0,5	3	8	4,5
	x_xx	0	0	0	0	0,5	3	9,5	3
	x_x_x	0	0	0	0	0	0	0	0
Nr. Aisle Interferences (within group)									
	xy_sr_ss	5	7	9	11	4,88406	5,95652	4,92157	2,1044
	xy_sr_ds	8	11	14	17	7,85507	9,9565	9,7451	6,7307
	xy_hr	67	64	61	58	67,1304	65,0435	65,1667	67,582
Nr. Aisle Interferences (between group)									
	xy_sr_ss	0	0	0	0	0,00084	0,01134	0,0594	0,0650
	xy_sr_ds	0	0	0	0	0,00105	0,01134	0,06110	0,06632
	xy_hr	1	1	1	1	1	1,01512	1,20531	2,16444
T. Seat Interferences		**72**	**72**	**72**	**72**	**70**	**60**	**37**	**10,5**
T. Aisle Interferences		**81**	**83**	**85**	**87**	**80,871**	**81,99**	**81,16**	**78,71**
Total Interferences		**153**	**155**	**157**	**159**	**150,87**	**141,99**	**118,15**	**89,21**

Table 4. Simulation result of model MINLP and BF using ProModel

	BF3	BF4	BF5	BF6
Avg. Seat Interferences	70,76	72,11	73,36	72,22
Avg. Aisle Interferences	53,41	53,36	52,74	52,27
Avg. Total	124,17	125,47	126,1	124,49
Avg. Boarding Time	1436,76	1460,68	1473,69	1491,68
	MINLP3	**MINLP4**	**MINLP5**	**MINLP6**
Avg. Seat Interferences	70,95	59,38	36,96	10,46
Avg. Aisle Interferences	52,8	51,89	49,1	42,94
Avg. Total	123,75	111,27	86,06	53,4
Avg. Boarding Time	1431,02	1434,66	1430,6	1389,89

11. CONCLUSION

1. Mixed Integer Nonlinear Programming can be used to produce a model boarding.
2. Analysis calculation results show that the MINLP 6 model is 43.89% better than BF 6.
3. Promodel simulation results for sum of interference shows that MINLP 6 model is 57.1% better than BF 6 model.
4. For boarding time, MINLP 6 model is 6.82% better than the BF 6 model

Further Research

1. Boarding model can be formulating using quadratic.
2. It would be better if this new boarding model can be implemented in real aircraft.

12. REFERENCES

[1]. Bazargam M.2006. A linier programming approach for aircraft boarding strategy, European Journal of Operational Research 183, 394-411.

[2]. Borchers, Brian and Mitchell, J. E., September 1991. "An improved branch and bound algorithm for mixed integer nonlinear programs". R.P.I Math Report No. 200.

[3]. Bussieck, Michael R., 2003. Mixed-Integer Nonlinear Programming. GAMS Development Corporation, Washington DC.

[4]. Dolan, Elizabeth D, et all., 2002, The NEOS Server Optimization Version 4 and Beyond, Argonne National Laboratory.

[5]. Law, Averill M., 2007. Simulation Modelling and Analysis: McGraw Hil

[6]. Taha, Hamdy A. 2007. Operations Research: An Introduction. Toronto: Pearson Eduction, Inc.

[7]. Van Den Briel, M.H.L., Villalobos, J.R., Hogg, G.L., Lindemann, T., Mule, A.V., 2005. America west airlines develops eficient boarding strategies. Interfaces 35, 191–201.

[8]. Van Landeghem, H., Beuselinck, A., 2002. Reducing passenger boarding times in airplanes: A simulation based approach. European Journal of Operational Research 142, 294–308.

Optimizing Rijndael Cipher Using Selected Variants of GF Arithmetic Operators

Petrus Mursanto
Fakultas Ilmu Komputer Universitas Indonesia
Kampus UI, Depok 16424
Indonesia
Phone: +62-21-7863419
santo@cs.ui.ac.id

ABSTRACT
A series of experiments has been conducted to show that efficiency improvement in Galois Field (GF) operators does not directly correspond to the system performance at application level. The experiments were motivated by so many research works that focused on improving performance of GF operators. Numerous variants of operators were formed based on various combination of operation types (multiplication, division, inverse, square), representation basis (Polynomial, Normal, Dual), and processing types (serial, parallel). Each of the variants has the most efficient forms in either time (fastest) or space (smallest occupied area) when implemented in an FPGA chip. In fact, GF operators are not utilized individually, rather integrated one to the others to implement algorithms, mostly in cryptography and error correction applications. The experiments based on the implementation of Rijndael Cipher 128-bit using VHDL by means of two synthesis tools: the Xilinx ISE 8.2i and the Altium ProChip Designer concludes that application performance mainly depends on the composition and distribution of the operators as well as their interaction and interconnection within the system architecture.

Keywords
Galois Field, Rijndael, VHDL, FPGA.

1. INTRODUCTION
Galois Field (GF) arithmetic plays an important role in modern communication system, particularly in two important aspects of information exchange, i.e. security and data correctness. GF is utilized in cryptography algorithm [1][2] and error correction codes (ECC) [3][4]. Performance of applications in these two fields is determined by the efficiency of GF arithmetic operators involved in the system [5]. There has been found in the literatures research efforts in improving GF operators' efficiency, e.g. multiplication [6], division [7] and inversion [8]. In fact, GF operators are not performing their functions individually and independently, rather they are parts of a functional integration at system level. Is operator efficiency beneficial to the application level performance?

This paper reports an experimental result of implementing Rijndael encryption and decryption algorithms based on six variants of GF operator. The purpose of the experiment is to obtain an Rijndael configuration whose throughput is the most optimum. The Rijndael algorithm was implemented using VHDL by means of two synthesis tools: the **Xilinx** ISE 8.2i and the **Altium** ProChip Designer.

2. PREVIOUS RESEARCH
Similar to the ordinary algebra, GF algebra has a number of arithmetic operations, such as: addition, subtraction, multiplication, division, inversion, square and square root. Variants of GF arithmetic operators are characterized by:

1. operation types: multiplication, division, inversion, square or square root
2. representation basis: standard/polynomial (PB), normal (NB) or dual (DB)
3. processing types: serial or parallel

In digital circuit, GF addition and subtraction are simply implemented by exclusive-OR logic operation. The advance of digital technology has shifted performance measurement mechanism from the running time of software algorithm [9] to VLSI complexity, i.e. the number of components and their total delay [6].

The first circuit structure of GF arithmetic was proposed by Berlekamp in 1982, i.e. polynomial and dual based multiplication [10]. Normal based multiplier was introduced firstly by Massey-Omura in 1986 [11], which is known afterward as MO multiplier. In 1988, Mastrovito proposed a more modular multiplier with higher regularity of the structure that suits systolic cells in VLSI [12]. However, speed, size and modularity of Mastrovito's multiplier depend much on the irreducible polynomial $P(x)$ used to generate the field elements. By selecting the right $P(x)$, parallel multiplication has at most $2m^2-1$ gates and occupies 55% of the space required for implementing Bartee and Schneider's algorithm [9]. In 1991, Mastrovito's dissertation reported an experimental investigation on multiplication using more than one representation basis [13]. It was concluded that PB multiplier is the most versatile form for the most arithmetic computational problems $GF(2^m)$ in VLSI. In addition, PB solution also posses conversion cost that can compensate the efficiency gained by the other representation basis [14] and occupies a half space of the one required by MO multiplier. Mapping problem for interbasis conversion is the concern of Wu et al. [15] which introduced an efficient conversion method from PB to NB specifically for squaring. Furthermore,

Sunar et al. proposed conversion matrix for any form of generator polynomial [16].

Several improvements of multiplication algorithm were also reported by Afanasyev [17] and similarly by Hasan et al. [18] that proposed a modification of the architecture by defining the irreducible polynomial as all-one polynomial (AOP). By applying the AOP, Hasan claimed the complexity of multiplication decreases by 50%. Meanwhile, Lee-Lim also reported a performance improvement by applying circular dual basis (CDB) [19]. Lee's method is very efficient for trinomial with composite $GF((2^n)^m)$ where m is primary relative over n, or $gcd(m,n) = 1$. However, defining certain form of irreducible polynomial is considered as limitation, inflexible and low reusability [20].

A comprehensive study on GF arithmetics was reported by Paar's dissertation [21], in which he proposed a decomposition algorithm from $GF(2^k)$ to $GF((2^n)^m)$ where $k = n.m$, called **composite field**. In addition, Paar also explored inversion after the first algorithm introduced by Itoh-Tsujii in 1988 [22]. Further Paar's research in [23] reported composite field multiplication and inversion in $GF(2^8)$. Composite field implementation in FGPA showed component saving by 25% and acceleration by 10% [24]. The composite field inversion requires 29% of AND and XOR gates compared to the standard one. Rudra [25] and Jutla [26] also developed a method for linear transformation of GF binary elements to composite field representation.

Combination of serial dan parallel processes were reported by Choi et al. [27] that introduced hybrid multiplier by forming irreducible polynomial $x^m + x^n + 1$ where $n \leq m/2$. This hybrid multiplier has flexible structure to compromise space and time complexity and is proven having less complexity than Wu and Hasan multiplier [15]. Several other methods were proposed to support hardware implementation, such as Huang-Wu [28] that has systolic array architecture approach to ease the testing process.

Previous implementation of Rijndael cipher has been reported, such as improvement of arithmetics efficiency based on composite field by Rudra et.al. [25], optimization of transformation using Look Up Table by Lee [29][30], and performance improvement of SubBytes algorithm specifically on S-Box module by Rijmen [31].

3. MOTIVATION

Previous research focused on efficiency improvement of GF operators to obtain better performance in term of speed or occupied space when implemented in digital circuits. However, literatures on GF-based implementation at system level are merely experience sharing with specific features without any analysis on the consequences of GF operator variants involved in the system. Based on the available literatures, the experiments were designed to answer the following research questions:

1. Can performance improvement gained at operator level be obtained linearly at the application level?
2. How can GF-based circuit optimization be achieved at application level?
3. Will an application take benefit by employing all best variants of GF operators?

4. METHODOLOGY

A set of experiments was designed to examine whether the best variants of GF operator can be combined to construct the most efficient application. In other words, what configuration of GF operator variants can produce an application with the greatest throughput. Arithmetic operator types were taken as suggested by the algorithm of Rijndael cipher 128-bit. The operators were implemented as the most efficient variant from each combination of two parameters: representation basis (PB, NB or DB), and processing structure (parallel or serial). For further reference in the next discussion, we use six variants of operators as follows:

Table 1. GF operator variants

Variant	Structure	Basis
1	Parallel	Polynomial
2	Serial	
3	Parallel	Normal
4	Serial	
5	Parallel	Dual
6	Serial	

The Rijndael cipher is implemented into six versions, each of which was constructed based on the variants in Table 1. They were implemented using structural VHDL and synthesized with Xilinx ISE 8.2i and Altium ProChip Designer. The synthesis process results in maximum combinational path or total delay that defines the maximum frequency possibly supplied to the system. Hence, the system throughput can be calculated based on the data capacity proceeded per time unit, expressed in Mega Byte per second (MBps). Throughput is then used as an indicator of the system performance. An optimal configuration is defined as the one having biggest throughput among the six versions of the system.

5. GF OPERATOR ARCHITECTURES

This section briefly discusses the six variants of GF operator, in particular the one widely used in encryption and decryption processes, i.e. the multiplication. Variant 1 of multiplier implements Mastrovito's circuit in [12]. Multiplication is proceeded partially in variant 2 by the circuit shown in Figure 1. Variant 3 and 4 are realized based on the NB multiplier in [13]. Multiplier variant 5 and 6 are implemented using parallel and serial structure in [13].

The implementation of six variants of multiplier results in delays shown in Table 2. It can be seen that the parallel multiplier in dual basis has the biggest combination delay. The best variant is the one having the smallest delay, i.e. polynomial based parallel multiplier or variant 1.

6. RIJNDAEL CIPHER

This session describes the implementation of Rijndael cipher with focus on hardware structure involving GF algebra. Decipher is a reverse process of cipher and vice versa. More details on analytic theory and algorithm's philosophic background, its strength and weakness as well as limitation are covered in [32].

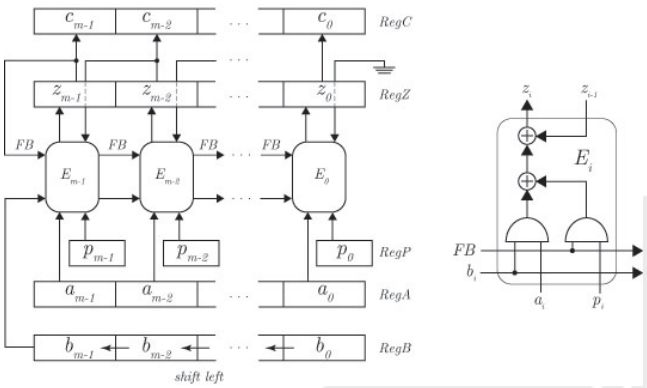


Figure 1. Serial PB multiplier *GF(2^m)*

Table 2. Delay of *GF(2^4)* multiplier in ns

Structure	Tools	PB	NB	DB
Parallel	Xilinx	12.69	13.49	17.61
	Altium	11.37	11.94	14.25
Serial	Xilinx	15.96	15.18	15.31
	Altium	15.24	16.26	15.14

According to Rijndael encryption scheme, one block of data is converted to ciphertext by means of a number of transformation algorithms. A temporary result of internal process in delivering ciphertext is called *State*. State is represented in a square of byte arrays. It has four rows and a number of columns. Number of columns is defined using a variable that equals to block length divided by 32. This session describes the implementation of Rijndael cipher with data block and key length of 128 bits.

Encryption algorithm is arranged in several steps of computation such as shown in Figure 2. In initial phase (iteration 0), State is computed by XORing data block and the cipher key. Furthermore, State is going through 10 iterations consisting of computation phase SubBytes (SB), ShiftRows (SR), MixColumns (MC) and AddRoundKey (ARK). Specific for the 10th iteration, State skips the MC process. In each iteration, result of MC is XORed with Round Key which is unique for corresponding iteration. The Round Key is produced by Round Key Generator from a number of transformations over the cipher key.

Detail of internal process in each block within Figure 2 has been discussed in [31]. This paper presents the structure of Rijndael's blocks and their throughput as consequences of GF operator variants selected in the implementations.

6.1 SubBytes

SubByte (SB) transformation is a byte non-linear substitution, operates on each State independently. Substitution table (known as S-Box) is invertible and built with composition of two main steps, i.e.: multiplicative inversion and Affine transformation. *GF(2^8)* inversion can be accomplished in either parallel or serial. Direct parallel inversion for *GF(2^4)* has been discussed in [12] with subfield in [13]. Affine transformation delay is duration time required by XOR operation. Considering additional cycles for Affine transformation, throughput of SubBytes is obtained as shown in Table 3.

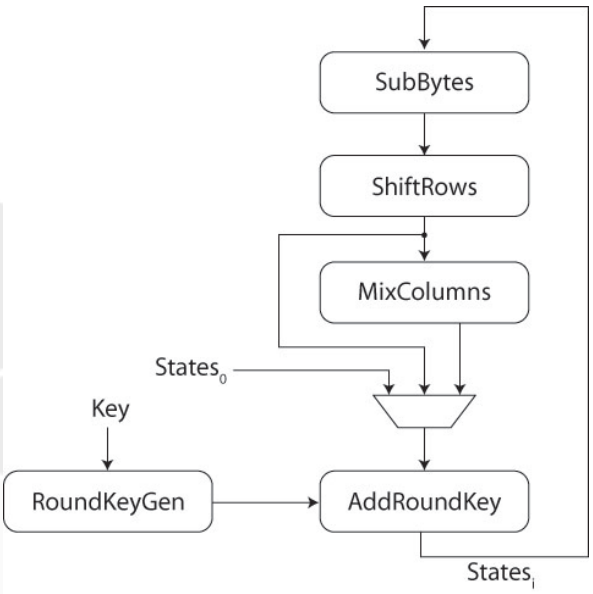


Figure 2. Rijndael Cipher process

Table 3. Throughput SubBytes in MBps

Structure	Tool	PB	NB	DB
FullPar	Xilinx	105.56	92.00	80.88
	Altium	113.56	98.40	86.36
Parallel	Xilinx	102.40	149.20	n.a.
	Altium	112.00	162.80	n.a.
Serial	Xilinx	116.60	131.40	120.48
	Altium	125.60	135.08	130.68

6.2 ShiftRows

In ShiftRows (SR), the row of State is shifted in *cyclic* or left rotated with offset varies from 0 to 3. SR is directing input bytes to different row of output bytes. Implementation delay equals to zero or 'cost-free' since there is no gate involved.

6.3 MixColumns

In MixColumn, column of State is considered as a polynomial of *GF(2^8)* and multiplied by modulo of x^4+1 with specific polynomial:

$$c(x) = \text{'03'}x^3 + \text{'01'}x^2 + \text{'01'}x + \text{'02'}$$

MixColumns delay is duration time of multiplication which is implemented in parallel or serial. In this case, there are multiplication with constant operands, i.e. 01, 02 and 03. Performance evaluation is conducted over the throughput of MixColumns if the multiplier is implemented in one of the six variants in Table 1. Result of throughput measurement using Xilinx and Altium tools is shown in Table 4.

Table 4. MixColumns throughput in MBps

Structure	Tool	PB	NB	DB
Parallel	Xilinx	819.6	746.8	620.0
	Altium	897.2	814.0	669.6
Serial	Xilinx	842.0	554.8	858.4

	Altium	907.2	570.4	931.2

6.4 AddRoundKey

This is a simple process, i.e. XORing State with the Round Key resulted from RKG in corresponding iteration. Delay of this process comes from the XOR gates.

6.5 Round Key Generator

RKG is to produce Round Key for each iteration. The original cipher key is only for iteration 0, and then used by RKG for delivering Round Key in iteration 1. Round Key 1 is processed for delivering Round Key in iteration 2, and so on. Circuit delay is the duration time required for multiplicative inversion. Round Key Generator has similar structure with SubBytes involving only inversion and XOR. Performance measurement of ARK can be seen in Table 5. It is shown that normal based inversion variant has the highest performance as demonstrated in [33].

Table 5. Throughput Round Key Generator in MBps

Structure	Tool	PB	NB	DB
FulPar	Xilinx	211.12	183.96	161.84
	Altium	227.16	196.80	172.72
Parallel	Xilinx	117.08	186.76	n.a.
	Altium	128.20	203.44	n.a.
Serial	Xilinx	118.44	134.96	122.64
	Altium	127.56	138.72	133.00

6.6 Rijndael Cipher Performance

The overall performance of Rijndael encryption requires a uniform of symbol representation basis. For that reason, the best performance must be identified for every representation basis. As summary, Table 6, Table 7 and Table 8 show the best throughput of the Rinjdael's modules, each for the three representation basis.

Table 6. The best throughput of SubBytes

Tool	Modul Architecture	Thrghput	Δ×clk
Xilinx	SymbSerial Serial PB	116.60	2.11×65
	SymbSerial Parallel NB	149.20	21.42×5
	SymbSerial Serial DB	120.48	2.33×57
Altium	SymbSerial Serial PB	125.60	1.96×65
	SymbSerial Parallel NB	162.80	19.66×5
	SymbSerial Serial DB	130.68	2.15×57

Table 7. The best throughput of MixColumns

Tool	Modul Architecture	Thrghput	Δ×clk
Xilinx	Serial Multp PB	842. 0	2.11×9
	Parallel Multp NB	746.8	21.42×1
	Serial DB	858.4	2.33×8
Altium	Serial Multp PB	907.2	1.96×9
	Parallel Multp NB	814.0	19.66×1
	Serial Multp DB	931.2	2.15×8

Table 8. The best throughput of Round Key Generator

Tool	Modul Architecture	Thrghput	Δ×clk
Xilinx	FullPar Inv PB	211. 12	75.79×1
	Parallel Inv NB	186.76	85.68×1
	FullPar Inv DB	161.84	98.87×1
Altium	FullPar Inv PB	227.16	70.44×1
	Parallel Inv NB	186.76	78.64×1
	FullPar Inv DB	172.72	92.64×1

Round Key Generator (RKG) runs in parallel with three iterative processes, they are: SubBytes (SB), ShiftRows (SR) and MixColumns (MC). Results of those parallel processes are XOR-ed with Add Round Key (ARK) to produce new State that becomes the input for next iteration. Hence, number of cycles required in one iteration is defined by the biggest one of the two processes whose period follows the slowest module. Throughput is calculated based on parallel processes in four columns; in this case the total data is 16 bytes.

Table 9. The highest throughput of encryption

Tool	Basis	Δ/clk	#total clk	Δ(ns)	Thrghpt (MBps)
Xilinx	PB	75.79	65+9+1	5684.25	2.82
	NB	85.68	5+1+1	599.76	26.68
	DB	98.87	57+8+1	6525.42	2.45
Altium	PB	70.44	65+9+1	5283	3.03
	NB	78.64	5+1+1	550.48	29.07
	DB	92.64	57+8+1	6114.24	2.62

It is shown in Table 9 that encryption performance is very low since it has to accommodate the delay of RKG which is built in parallel structure. It is therefore required to build a serial structure of RKG as long as the number of cycles in total does not exceed the number of cycles accumulated by SB, SR and MC. PB and DB can have operational structures in serial per symbol with serial operators. Whereas NB should have serial structure per symbol with parallel operators. In this case, serial operator results in number of cycles exceeds the number of cycles for encryption. Performance of serial structured RKG is presented by Table 10.

Table 10. Throughput Round Key Generator (RKG) serial

Tool	Modul Architecture	Thrghput	Δ×clk
Xilinx	Serial-Serial Inv PB	29. 61	2.11×64
	Serial-Parallel Inv NB	46.69	21.42×4
	Serial-Serial Inv DB	30.66	2.33×56
Altium	Serial-Serial Inv PB	31.89	1.96×64
	Serial-Parallel Inv NB	50.86	19.66×4
	Serial-Serial Inv DB	33.25	2.15×56

By changing RKG structure from parallel to serial, performance of the system increases. It is evidently shown by the increasing throughput of the application significantly in Table 11. It can be seen that Rijndael cipher with normal based representation has the highest throughput. This fact is shown consistently by both Xilinx and Altium tools. It proves that the most efficient at PB operators does not deliver the most optimal Rijndael application. Performance degradation of RKG modularly in fact improves the throughput of integrated system in the application.

Table 11. Throughput of Rijndael cipher with serial RKG

Tool	Basis	Δ/clk	#total clk	Δ(ns)	Thrhpt (MBps)
Xilinx	PB	2.11	65+9+1	158.325	101.06
	NB	21.42	5+1+1	149.94	106.71
	DB	2.33	57+8+1	153.78	104.04
Altium	PB	1.96	65+9+1	147	108.84
	NB	19.66	5+1+1	137.62	116.26
	DB	2.15	57+8+1	141.77	112.86

7. OPTIMIZATION BY THE SYNTHESIS TOOL

Experiment results show that employing the best operators does not automatically produce the most efficient application. This phenomenon is shown consistently by Xilinx and Altium synthesis tools. Simple and common logic saying that parallel process is faster than serial does not hold. This is caused by the fact that the synthesis tool does some improvement or limited optimization to the VHDL structural design.

Multiplication and inversion are obviously the dominant process in Rijndael encryption. Examination on the results shows that multiplication with fixed bit operands is optimized by removing unnecessary components in the system. For specific configuration of Rijndael cipher, *P(x)* and *g(x)* are fixed during the system lifecycle. *g(x)* is one of the operand performing as *a(x)* in Figure 1. Fixed values of p_i and g_i cause the content of block E_i in Figure 1 can be simplified becoming the circuits shown in Figure 3. All combination of a_i and p_i result in internal block E_i without any AND gate. By omitting AND gates, Xilinx saves significantly the delay so that the minimum period is 2,5 ns. Hence, processing 4 bit requires 10 ns, which is faster than parallel multiplication delay 12.69 ns.

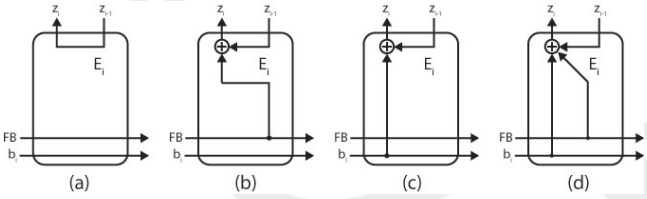


Figure 3. Internal component of E_i in PB serial multiplier for (a) a_i=0; p_i=0 (b) a_i=0; p_i=1 (c) a_i=1; p_i=0 (d) a_i=1; p_i=1

It is examined that optimization was not applied to parallel multiplications although they also have constant operands. It is due to the fact that the tools optimize constant bit only, whereas parallel multiplier signal is implemented as an *m*-bit bus.

DB multiplier is superior over the PB and NB variants in term of fully serial delivery of the product. With constant values of *P(x)* and *A(x)*, DB serial multiplier delivers the product bit by bit as the operand b_i enters from MSB to LSB. Therefore, variant 6 saves 1 cycle compared to variant 2 and 4 that delivering the product after the whole cycle for *m*-bit completes. For that reason, variant 6 based system can proceed addition serially as the product is delivered from index *m-1* to *0*.

With constant $P(x) = x^4 + x + 1$, optimization steps applied to serial DB multiplication is shown by Figure 4. In general this finding supports several statements in the literature that multiplication with constant operands can be more efficient [10]. It was examined as well that optimization does not apply to serial NB multiplier. In serial NB multiplication, the values of both operands change dynamically during the lifecycle of the system due to the rotation of internal registers.

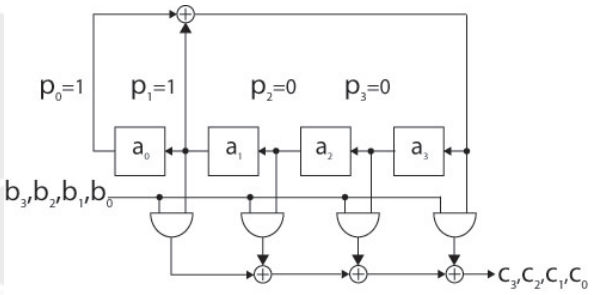


Figure 4. DB serial multiplier with constant *P(x)*

8. CONCLUSION

This paper reports that an optimal performance of Rijndael cipher does not always require the best variants or the most efficient GF operators. Combining all operators, each of which is the most efficient variants, is not a simple mechanistic conversion process. Obtaining synergic efficiency at system level requires careful considerations on several factors such as: the operator composition and distribution, interaction between them and types of internal process within the system. In addition, when implemented in FPGA, efficiency improvement is also contributed by synthesis tools that optimize serial operators whose operands are constant. This explains why the experimental results show the superiority of serial based system over parallel ones.

It is interesting to examine further the consistency of this optimization in other GF based applications, such as error correction codes. Explorative experiments are required for Rijndael AES with cipher-key 8-bit [34] or RS(255,223) 8-bit such as the one used by NASA [35]. Hypothetical prediction suggests that higher performance ratio would be obtained by serial variants over the parallel ones. It is because of additional combinational path in parallel operators that results in bigger delays. Meanwhile, serial operators requires only several additional cycles with constant minimum periods.

9. ACKNOWLEDGMENTS

Part of this material is based upon work supported by European Union Project VN/Asia-Link/001 (79754). The author would like to thank Prof. R.G. Spallek and Mr. Jörg Schneider for providing tools and facilities in the Professur für VLSI-Entwurfssysteme, Diagnostik und Architektur, Technische Universität Dresden, Germany.

10. REFERENCES

[1] Tilborg, H.v. 1988. *An Introduction to Cryptology*, Kluwer Academic Publishers.

[2] Schneier, B. 1993. *Applied Cryptography*. Wiley & Sons.

[3] B.Wicker, S. 1995. *Error Control Systems for Digital Communication and Storage*. Prentice Hall, Upper Saddle River, New Jersey 07458.

[4] Sweeney, P. 2002. *Error Control Coding from Theory to Practice*. John Wiley & Sons, Inc., England.

[5] Lin, S. and Costello, D. J. 1983. *Error Control Coding*. Prentice Hall, Englewood Cliffs, New Jersey.

[6] Wang, C., Truong, T., Shao, H., Deutsch, L., Omura, J., and Reed, I. Aug 1985. VLSI Architecture for Computing Multiplications and Inverses in *GF(2^m)*. *IEEE Trans. on Computers* 34(8), 709–717.

[7] Hasan, M., Wang, M., and Bhargava, V. Aug 1992. Division and Bit-Serial Multiplication over *GF(q^m)*. *IEEE Trans. on Computers* 41(8), 972–980.

[8] Zhou, T., Wu, X., Bai, G., and Chen, H. 4 Jul 2002. Fast GF(p) Modular Inversion Algorithm Suitable for VLSI Implementation. *Electronics Letters* 38(14), 706–707.

[9] Bartee, T. and Schneider, D. March 1963. Computation with Finite Fields. *Information and Computers* 6, 79 – 98.

[10] Berlekamp, E. Nov 1982. Bit-Serial Reed-Solomon Encoders. IEEE Trans. Information Theory 28, 869 – 874.

[11] Massey, J. and Omura, J. 1986. *Computational method and apparatus for finite field arithmetic*. Technical report US Patent No. 4,587,627 to OMNET Association Sunnyvale CA, Washington, D.C. Patent and Trademark Office.

[12] Mastrovito, E. D. Dec 1988. *VLSI designs for computations over finite fields GF(2^m)*. Master Thesis. Linkping Studies in Science and Technology Linkping, Sweden. Thesis No: 159.

[13] Mastrovito, E. 1991. *VLSI Architectures for Computations in Galois Fields*. PhD Dissertation. Department of Electrical Engineering, Linkping University, Sweden.

[14] Gollman, D. 1990. *Algorithmenentwurf in der kryptographie*. Habil. University of Karlsruhe, Preprint.

[15] Wu, H., Hasan, M. A., F.Blake, I., and Gao, S. 23 Aug 2001. *Finite Field Multiplier using Redundant Representation*.

[16] Sunar, B., Savas, E., and Koç, Ç.K. Nov 2003. Constructing Composite Field Representations for Efficient Conversion. *IEEE Trans. on Computers* 52(11), 1391–1398.

[17] Afanasyev, V. (1990) Complexity of VLSI Implementation of Finite Field Arithmetic. In *Proc. II Int'l Workshop on Algebraic and Combinatorial Coding Theory*, Leningrad, USSA. pp. 6 – 8.

[18] Hasan, M., Wang, M., and Bhargava, V. Oct 1993. Division and Bit-Serial Multiplication over GF(q^m). *IEEE Trans. on Computers* 42(10), 1278 – 1280.

[19] Lee, C.-H. and Lim, J.-I. 1998. A new aspect of dual basis for efficient field arithmetic. *Technical report Samsung Advanced Institute of Technology*.

[20] Fenn, T. S., Benaissa, M., and Taylor, D. March 1996. *GF(2^m)* Multiplication and Division Over the Dual Basis. *IEEE Trans. on Computers* 45(3), 319 – 327.

[21] Paar, C. Jun 1994. *Efficient VLSI Architectures for Bit-Parallel Computation in Galois Fields*. PhD Dissertation. Institute for Experimental Mathematics, University of Essen Essen, Germany.

[22] Itoh, T. and Tsujii, S. 1988. A Fast Algorithm for Computing Multiplicative Inverses in *GF(2^m)* using Normal Basis. *Information and Computing* 78, 171–177.

[23] Paar, C. and Rosner, M. Apr 1997. Comparison of Arithmetic Architectures for Reed-Solomon Decoders in Reconfigurable Hardware. *IEEE Symposium on FPGA-Based Custom Computing Machines* (FCCM) pp. 219–225.

[24] Mursanto, P. 29-30 Jan 2007. Comparison of GF Multipliers in Standard and Composite Field Architecture. In *National Conference on Computer Science & Information Technology*. Depok, Fasilkom UI .

[25] Rudra, A., Dubey, P. K., Jutla, C. S., Kumar, V., Rao, J., and Rohatgi, P. 2001 vol.2162, Efficient Rijndael Encryption Implementation with Composite Field Arithmetic. In *Proc. 3rd Int'l Workshop on Cryptographic Hardware and Embedded Systems*, London, UK. pp. 171–184.

[26] Jutla, C., Kumar, V., and Rudra, A. On the circuit complexity of isomorphic Galois Field transformations *Technical Report* RC22652 (W0211-243) IBM Research Division, Nov 2002.

[27] Choi, Y., Chang, K.-Y., Hong, D., and Cho, H. 8 July 2004 Hybrid Multiplier for *GF(2^m)* Defined by Some Irreducible Trinomials. *Electronics Letters* 40 (14) 852 – 853.

[28] Huang, C.-T. and Wu, C.-W. Sep 2000. High-Speed Easily Testable GF Inverter. *IEEE Trans. Circuit and Systems* 48(9), 909–918.

[29] Li, H. 27-29 June 2004 A Parallel S-Box Architecture for AES Byte Substitution. *International Conference on Communications, Circuits and Systems* 1(1), 1 – 3.

[30] Li, H. 23-26 May 2005 A New CAM Based S-Box Look Up Table in AES. *IEEE International Symposium on Circuits and Systems* 5, 4634–4636.

[31] Rijmen, V. 2001. *Efficient implementation of the rijndael S-box* F.W.O. Postdoctoral Report, sponsored by the Fund for Scientific Research – Flanders Belgium Katholieke Universiteit Leuven, Dept. ESAT, Belgium.

[32] Daemen, J. and V., R. *AES Proposal: Rijndael* March 9, 1999 Version 2.

[33] Mursanto, P. 19 July 2007. Manfaat Representasi Elemen Berbasis Normal dalam Meningkatkan Kinerja Operator Aritmetika Galois Field. In *Proc. 6th National Conference Design and Application of Technology Faculty of Engineering*, Widya Mandala Catholic University Surabaya : pp. 121–127.

[34] Daemen, J. and Rijmen, V. March 2001. *Rijndael, The Advanced Encryption Standard*. Dr.Dobb's Journal 26, 137–139.

[35] Sklar, B. 2003. *Digital Communications Fundamentals and Applications*, Prentice Hall, Inc., New Jersey 2nd ed.

PCR Primer Design Using Particle Swarm Optimization Combined with Piecewise Linear Chaotic Map

Cheng-Hong Yang
Department of Electronic Engineering, National Kaohsiung University of Applied Sciences
415 Chien Kung Rd., Kaohsiung 807, Taiwan, R.O.C.
+886-7-3814526 ext. 5639
chyang@cc.kuas.edu.tw

Yu-Huei Cheng
Department of Electronic Engineering, National Kaohsiung University of Applied Sciences
415 Chien Kung Rd., Kaohsiung 807, Taiwan, R.O.C.
+886-7-3814526 ext. 5639
yuhuei.cheng@gmail.com

Li-Yeh Chuang
Department of Chemical Engineering & Institute of Biotechnology and Chemical Engineering, I-Shou University
No.1, Sec. 1, Syuecheng Rd., Dashu Township, Kaohsiung 840, Taiwan, R.O.C.
+886-7-6577711 ext. 3421
chuang@isu.edu.tw

ABSTRACT

Many methods of primer design have been proposed to provide feasible primer sets for polymerase chain reaction (PCR) experiments. However, most of these methods is time consuming to design optimal primers from large quantities of template DNA, and they are usually fail to provide a specific size of PCR product. Particle swarm optimization (PSO) has been applied to solve all kinds of problems and proved to be effective. In this paper, a piecewise linear chaotic map (PWLCM) is proposed to determine the value of inertia weight of PSO (PWLCPSO) to design feasible primers. The primer sets for Homo sapiens RNA binding motif protein 11 (RBM11), mRNA (NM_144770), and Homo sapiens tripartite motif-containing 72 (TRIM72), mRNA (NM_001008274) were designed by PSO and PWLCPSO. Five hundred runs were performed on different PCR product lengths and the melting temperature was calculated by different methods. A comparison of the results obtained from PSO and PWLCPSO primer design showed that PWLCM provided better primer sets than PSO primer design.

Keywords

Polymerase chain reaction (PCR), primer design, particle swarm optimization (PSO), piecewise linear chaotic map (PWLCM).

1. INTRODUCTION

Polymerase chain reaction (PCR) is a common technology used for fast mass duplication of DNA sequences [Mullis and Faloona 1987]. It has been widely applied on the fields of biology and medicine. A feasible primer set is essential for PCR performance. However, it is a tedious work to design a feasible primer set from a template sequence manually. Numerous primer design constraints, such as the length, length difference, GC content, melting temperature (T_m), difference of melting temperature ($T_{m\text{-}diff}$), GC clamp, dimer, hairpin and specificity need to be considered that makes difficulty to find a feasible primer set artificially. Manually primer design is unsuitable due to the complexity of processes involved. Accordingly, it is preferable to design primer sets using automatic computation.

At present, many primer design methods have been proposed to design primer sets. Kämpke *et al.* implement dynamic programming [Kämpke *et al.*, 2001] to design primers. The advantageous of this method is able to provide multiple primers from multiple target DNA sequences. However, it takes a relatively long time to obtain a suitable primer set. Chen *et al.* based on thermodynamic theory to evaluate the fitness of primers and developed a succinct web-based tool called PDA for primer design [Chen *et al.*, 2003]. Wu *et al.* proposed a genetic algorithm (GA) to design optimal primer set imitating nature's process of evolution and genetic operations on chromosomes [Wu *et al.*, 2004]. Hsieh *et al.* used automatic variable fixing and redundant constraint elimination to tackle the binary integer programming problem associated with the minimal primer set (MPS) selection problem [Hsieh *et al.*, 2003]. Wang *et al.* employed a greedy algorithm to generate a MPS that is specifically annealed to all open reading frames (ORFs) in a given microbial genome to improve the hybridization signals of microarray experiments [Wang *et al.*, 2004]. Miura *et al.* identified the specificity-determining subsequence (SDSS) of each primer and examined its uniqueness in a target genome [Miura *et al.*, 2005]. In our previous study, we applied a memetic algorithm (MA) [Yang *et al.*, 2009] and particle swarm optimization (PSO) method [Yang *et al.*, 2010] to search for feasible primers.

PSO has been applied to all kinds of problems and yielded promising results during the past decade. PSO, developed by Kennedy and Eberhart in 1995 [Kennedy and Eberhart 1995], is a population-based stochastic optimization technique that simulates the social behavior of organisms, such as birds in a flock, and describes an automatically evolving system. However, many studies in the literature report the premature convergence of PSO and that the process may get trapped in a local optimum easily when handling complex multimodal problems [Hsieh *et al.*, 2009; Liang *et al.*, 2006; Van den Bergh and Engelbrecht 2004; Yang *et al.*, 2007]. Chaos is ergodic and stochastic, and contains elements of certainty. By following chaotic orbits, a global optimum or a good approximation may eventually be reached with high probability in a dynamic system. In this paper, we use a piecewise linear chaotic map (PWLCM) to determine the value of inertia weight of PSO, named PWLCPSO, to improve the performance of the primer design. Different PCR product lengths and melting temperature calculations are used to evaluate and compare the performances of PSO and PWLCPSO.

2. PROBLEM DESCRIPTION

This section describes the primer design problem. Let T_D be the template DNA sequence, which is made up of base-nucleic acid codes of the DNA, i.e., 'A', 'T', 'C', or 'G'. T_D can then be defined as follows:

$$T_D = \{B_i \mid \forall B_i \in \text{'A' or 'T' or 'C' or 'G'}, i \in Z^+\} \quad (1)$$

where, B represents the base-nucleic acid sequence made up of the base-nucleic acid codes of the DNA; i is the index of the position on T_D, and Z^+ represents the region of positive integers.

The primer design problem now consists of finding a pair of sub-sequences in T_D for satisfying various constraints. One sub-sequence is called the forward primer and the other is called the reverse primer. The forward primer and the reverse primer are defined as follows:

$$P_f = \{B_i \mid \forall B_i \in \{'A','T','C','G'\}, F_s \le i \le F_e \le T_D, i \in Z^+\} \quad (2)$$

$$P_r = \{\overline{B_i} \mid \forall B_i \in \{'A','T','C','G'\}, R_s \le i \le R_e \le T_D, i \in Z^+\} \quad (3)$$

where, P_f is the forward primer, and F_s and F_e denote the start index and the end index of P_f in T_D. P_r is the reverse primer, and R_s and R_e denote the start index and the end index of P_r in T_D. Together, P_f and P_r are called a primer pair. The anti-sense sequence of B is called $\overline{B}$. This anti-sense sequence $\overline{B}$ is the reverse of the complementing sequence of B.

In Fig. 1, the symbols are described as the length of the template DNA is T_l, the minimum PCR product length is P_{min}, the maximum PCR product length is P_{max}, the start position of the forward primer is F_s, the length of the forward primer is F_l, the PCR product length between the forward primer and the reverse primer is P_l, the length of the reverse primer is R_l, the random range of F_s is F_{s_Range}, and the length from F_s to the template DNA end is P_{Range}. In order to determine a primer pair, a vector given by F_s, F_l, P_l and R_l is used. We define this vector P_v as:

$$P_v = (F_s, F_l, P_l, R_l) \quad (4)$$

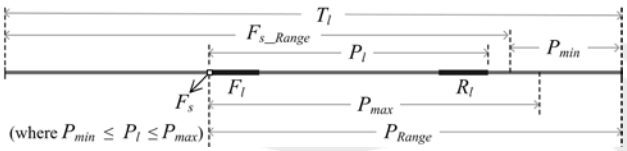


Figure 1. Parameters of the template DNA and primer set.

Table 1. parameters used in Figure 1.

Parameter	Description
F_s	Start position of the forward primer
F_l	Length of the forward primer
P_l	PCR product length between forward primer and reverse primer
R_l	Length of the reverse primer
F_{s_Range}	Random range of F_s
P_{min}	Minimum PCR product length
P_{max}	Maximum PCR product length
P_{Range}	Length from F_s to the template DNA end
T_l	Length of template DNA

The reverse primer start index can obtain by calculating the following equation:

$$R_s = F_s + P_l - R_l \quad (5)$$

Therefore, the forward primer (P_f) and the reverse primer (P_r) can both be obtained through P_v. This vector P_v is the prototype of a particle used in the PSO, and later sections will employ P_v to perform primer design. Table 1 summarizes the parameters used in Fig. 1.

3. PRIMER DESIGN METHOD

The flowchart of the proposed method is shown in Fig. 2. Seven separated processes of 1) initialization of particle swarm; 2) determination of initial inertia weight; 3) evaluation of fitness value; 4) judgment of termination condition; 5) finding *pbest* and *gbest*; 6) updating of inertia weight using PWLCM; and 7) updating of velocity and position of each particle, are described below.

3.1 Initialization of particle swarm

Initially, ten particles $P_v = (F_s, F_l, P_l, R_l)$ are randomly generated as an initial particle swarm without duplicates. F_s is randomly generated between 1 and $(T_l - P_{min} + 1)$. F_l is randomly generated between the minimum length of the primer and the maximum length of the primer. In the present study, the minimum length and the maximum length of the primer was set to 16 bps and 28 bps, respectively. In order to limit the PCR product length, a random P_l is generated between P_{min} and P_{max}. R_l was randomly generated in the same way as F_l. Each particle is given a velocity (v). This velocity is randomly generated within 0~1.

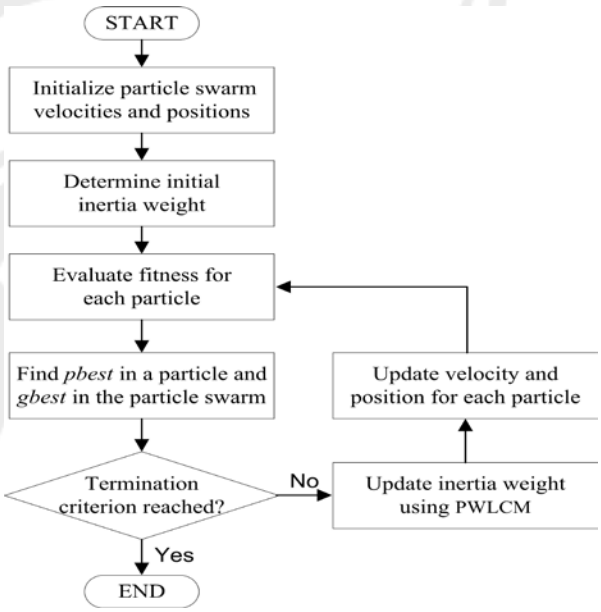


Figure 2. Flowchart of the proposed algorithm for PCR primer design.

3.2 Determination of initial inertia weight

Since chaos is highly sensitive to its initial conditions, small differences in the initial conditions yield widely diverging outcomes and can make long-term prediction impossible [Kellert 1993]. This inertia weight is subsequently updated and fine tuned by a chaotic system. In this study, an initial inertia weight of 0.8 which is commonly used in PSO to test the performance of PWLCPSO.

3.3 Evaluation of fitness value

The fitness function is an important core used to evaluate the fitness value of each particle in order to check whether the candidate primers satisfy the design constraints or not. Various primer design constraints are used as estimated values of the fitness function, and the produced fitness value is minimized.

For a longer primer, its specificity may be better, and a relatively high T_m is required. On the other hand, for a relatively short primer which specificity might be worse. Hence, neither a primer that is too long nor too short is suitable. A primer length between 16 bps and 28 bps is considered feasible for a PCR experiment [Wu, Lee, Wu and Shiue 2004]. In this study, the primer length constraint is not included in the fitness function, because F_l and R_l are always limited between the minimum length and the maximum length of the primer under the setting constraint conditions. The fitness value is provided by the following fitness functions, which are made up of $Len_{diff}(P_v)$, $T_m(P_v)$, $T_{mdiff}(P_v)$, $GC_{proportion}(P_v)$, $GC_{clamp}(P_v)$, $dimer(P_v)$, $hairpin(P_v)$ and $specificity(P_v)$, each of which is described below:

$$\begin{aligned} Fitness(P_v) = {} & 3 * (Len_{diff}(P_v) + GC_{proportion}(P_v) + GC_{clamp}(P_v)) \\ & + 10 * (Tm(P_v) + Tm_{diff}(P_v) + dimer(P_v) \\ & + hairpin(P_v)) + 50 * specificity(P_v) \end{aligned} \tag{6}$$

The weights of components in the fitness function are 3, 10 and 50 based on their importance. A larger weight indicates that the constraint is more important, and *vice versa*. The weights can be adapted freely by users according to the different experimental conditions.

In the different primer length, less than or equal to 3 bps of difference between the forward primer and the reverse primer is considered optimal [Wu, Lee, Wu and Shiue 2004]. The $Len_{diff}(P_v)$ function is used to check this condition. An appropriate T_m in a primer is experimental range of 50-62°C. The $Tm(P_v)$ function is used to check whether the melting temperature of a primer pair is between 50°C and 62°C. The $Tm_{diff}(P_v)$ function is used to check whether the T_m difference between the forward and reverse primer exceeds 5°C. A lower T_m difference indicates a better primer pair. In this study, we use Wallace formula [Wallace *et al.*, 1979] to calculate the melting temperatures (T_m) of primers. The computational formula for T_m is:

$$Tm_W(P) = (\#G + \#C) * 4 + (\#A + \#T) * 2 \tag{7}$$

where, P represents the forward primer or reverse primer, #G represents the number of 'G', #C represents the number of 'C', #A represents the number of 'A' and # T represents the number of 'T'. The suffix W represents the formula which was proposed by Wallace.

Furthermore, we also use a more elaborate equation proposed by Bolton and McCarthy [Sambrook *et al.*, 1989] to calculate the melting temperatures (T_m) of primers. The equation takes the ionic strength, G and C content and the length of the primer into account below.

$$\begin{aligned} Tm_{BM}(P) = {} & 81.5 + 16.6(\log_{10}[Na^+]) \\ & + 0.41 * (\text{GC content}) - 675/|P| \end{aligned} \tag{8}$$

where, P represents a primer and $|P|$ represents the length of primer P; $[Na^+]$ is the molar salt concentration. The suffix BM represents the formula which was proposed by Bolton and McCarthy.

An appropriate GC proportion in a primer should be in the range of 40-60%. The $GC_{proportion}(P_v)$ function calculates the ratio of nucleotide G and C that evaluates the GC proportion in a primer. In order to ensure a designed primer has a tightly localized hybridization bond, the $GC_{clamp}(P_v)$ function is used to check whether the 3' terminal end of a primer is G or C. Furthermore, the $dimer(P_v)$ function is used to check whether the forward primer and the reverse primer anneal to each other or anneal to themselves. The $hairpin(P_v)$ function is used to check if a primer anneal to itself. The annealing of primers is detrimental to the PCR experiment. Finally, the $specificity(P_v)$ function is used to judge whether the primer reappears itself in the template DNA sequence, and thus it ensures the specificity of the primer. The PCR experiment is more easily successful if the primer is specific which means it is annealed to a specific position within a template sequence.

3.4 Judgment of termination condition

When *gbest* has achieved the best position, the proposed method is terminated, i.e., its fitness value is 0, or when a maximum number of generations have been reached. When the termination condition is reached, *gbest* is the optimal solution of the primer design.

3.5 Finding *pbest* and *gbest*

In PSO, each particle has a memory of its own best experience. This is true for PWLCPSO as well. Each particle needs to find its personal best position and velocity (called *pbest*), and all particles must determine the global best position and velocity (called *gbest*). If the fitness of a particle P_v in the current generation is better than the fitness of *pbest* in the previous generation, *pbest* will be updated to P_v in the current generation. If the fitness of a particle P_v is better than *gbest* in the previous generation and is the best one in the current generation, *gbest* will be updated to P_v. Based on *pbest* and *gbest*, each particle adjusts its direction and moves to close the target in the next generation.

3.6 Updating of inertia weight using PWLCM

The inertia weight of PSO is used to balance the global and local search ability. A large inertia weight facilitates a global search while a small inertia weight facilitates a local search [Shi *et al.*, 2001]. In order to adjust the search ability, the inertia weight is changed dynamically using a chaotic system. In this paper, a

piecewise linear chaotic map (PWLCM) [Xiang *et al.*, 2007] is used to generate chaotic sequence for updating inertia weight. The chaotic map determines the value of inertia weight described below.

$$w(t+1) = \begin{cases} w(t)/p & , w(t) \in (0, p) \\ (1-w(t))/(1-p) & , w(t) \in [p, 1) \end{cases} \quad (9)$$

where the value of w at (t+1)th iteration is represented by $w(t+1)$; p is set as 0.7.

3.7 Updating of velocity and position of each particle

In each generation, all particles change their position and velocity. Equations (20) and (21) give the updating formulas for each particle.

$$v_i^{next} = w \times v_i^{current} + c_1 \times r_1 \times (s_i^p - s_i^{current}) + c_2 \times r_2 \times (s_i^g - s_i^{current}) \quad (10)$$

$$s_i^{next} = s_i^{current} + v_i^{next} \quad (11)$$

In equations (10) and (11), v_i^{next} is the updated velocity of the ith particle; $v_i^{current}$ is the current velocity of the ith particle; c_1 and c_2 are the acceleration coefficients; w is the inertia weight; r_1 and r_2 is a uniform random value which is randomly generated within 0~1; s_i^p is the personal best position of the ith particle; s_i^g is the global best position of the particles; $s_i^{current}$ is the current position of the ith particle; s_i^{next} is the updated position of the ith particle. When a particle is overshooting the ranges of F_s, F_l, P_l and R_l after updating, a preventive mechanism which uses a random process to reset their correct position is performed.

4. RESULTS AND DISSCUSSIONS

4.1 Data sets and environment

Two template sequences of Homo sapiens RNA binding motif protein 11 (RBM11), mRNA (NM_144770), and Homo sapiens tripartite motif-containing 72 (TRIM72), mRNA (NM_001008274) were tested using PSO and the proposed method PWLCPSO for primer design. Five main parameters, namely the number of iterations (generations), the number of particles, the inertia weight w, and the acceleration coefficient c_1 and c_2 were set in the PSO and PWLCPSO primer design methods for the computational simulations. These respective values were 100, 10, 0.8, 2 and 2. Five hundred runs in total were performed using the PSO and PWLCPSO primer design methods, with PCR product lengths in 150~300 bps, 500~800 bps and 800~1000 bps, and T_m calculated by the Wallace formula and the Bolton and McCarthy formula. The *in silico* simulated environment used a Pentium 4 CPU 3.4 GHz and 1GB of RAM under Microsoft Windows XP SP3.

4.2 Comparison of the primer design results

The results obtained from PSO and PWLCPSO primer design methods for Homo sapiens RNA binding motif protein 11 (RBM11), mRNA (NM_144770), and Homo sapiens tripartite motif-containing 72 (TRIM72), mRNA (NM_001008274) with different product lengths and T_m calculations are shown in Table 2 and Table 3, respectively.

The average accuracies of 78.2 % and 75.6% were reached when PWLCPSO primer design method with the Wallace formula was used for NM_144770 and NM_001008274 primer design with different product lengths. However, the average accuracies only got up to 67.5% and 61.1% when PSO was used under the same circumstances. The accuracies of PWLCPSO primer design method are thus 10.7% and 14.5% higher than for PSO primer design method for the two test template sequences. Furthermore, the average accuracies reached 72.7% and 65.6% when PWLCPSO was used with T_m calculation by the Bolton and McCarthy formula. The average accuracies only got up to 57.3% and 39.8% when PSO was used under these conditions. The accuracies of PWLCPSO primer design method were thus 15.4% and 25.8% higher than the accuracies of PSO primer design method for both test template sequences.

The PWLCPSO primer design method also outperformed the PSO primer design method in terms of the average running time with different T_m calculations. The computationally simulated results show that the performance of the proposed PWLCPSO method is superior to the performance of PSO on the primer design problem.

4.3 The effect of PWLCM for inertia weight

Chaos is a deterministic, random process found in non-linear system with a greatly sensitive to its initial conditions. Small differences in initial conditions yield widely diverging outcomes making long-term prediction impossible [Kellert 1993]. Mathematically, chaos may be considered a source of randomness since its simple deterministic dynamical behavior. PWLCM is a chaotic map with a simplicity in representation, efficiency in implementation, as well as good dynamical behavior [Xiang, Liao and Wong 2007]. A uniform invariant density function is known on its definition intervals [Baranovsky and Daems 1995]. In this study, the PWLCM is used to generate chaotic sequences for updating the inertia weight. Since PWLCM has widely range of parameter choices ergodic in (0, 1) with uniform distribution of w, it was introduced to control the movement of the particles. This may allow us to eventually reach a good approximation of the optimal results with high probability.

5. CONCLUSION

Primer design is an important issue in the related fields of molecular biology. The qualities of primers always influence PCR experiments. Although, many primer design methods and tools have been developed, most of them are inefficient falling short of the better qualities of primers. PSO is considered an efficient algorithm widely applied to solve various optimization problems. However, PSO tends to get trapped in a local optimum easily when applied to complex problems. In this study, PWLCM embedded in PSO is proposed to improve the performance of primer design.

The proposed PWLCPSO designs optimal primers with various primer constraints, such as primer length, primer length difference, GC proportion, PCR product length, melting temperature (T_m), melting temperature difference ($T_{m\text{-}diff}$), GC clamp, dimers (including cross-dimer and self-dimer), hairpin and specificity used

to appraise the fitness values. Each constraint was given a suitable weight based on its significance. Through the evolution of a fitness function, more feasible primer sets could always be obtained using PWLCPSO method than PSO. The proposed primer design method, PWLCPSO, could be a valuable tool for biologists and researchers involved in the related research fields.

6. ACKNOWLEDGMENTS

This work is partly supported by the National Science Council in Taiwan under grant NSC96-2221-E-214-050-MY3, NSC98-2221-E-151-040-, NSC98-2622-E-151-001-CC2 and NSC98-2622-E-151-024-CC3

Table 2. Accuracy and running time for PSO and PWLCPSO primer design methods. Computationally simulated results for Homo sapiens RNA binding motif protein 11 (RBM11), mRNA (NM_144770) using the Wallace formula and Bolton and McCarthy formula with PCR product lengths in 150 ~ 300 bps, 500~800 bps and 800~1000 bps. a, accuracy (%); t, running time (ms). Boldface indicates highest values.

T_m formula and primer design methods	Wallace's formula				Bolton and McCarthy formula			
	PSO		PWLCPSO		PSO		PWLCPSO	
PCR product length	a (%)	t (ms)	a (%)	t (ms)	a (%)	t (ms)	a (%)	t (ms)
150~300 bps	68.0	371562	80.0	309328	56.0	435235	74.0	364843
500~800 bps	69.6	387406	81.6	314250	58.4	424687	71.0	354219
800~1000bps	65.0	395454	73.0	347328	57.4	425094	73.0	347328
average	67.5	384807	78.2	323635	57.3	428339	72.7	355463

Table 3. Accuracy and running time for PSO and PWLCPSO primer design methods. Computationally simulated results for Homo sapiens tripartite motif-containing 72 (TRIM72), mRNA (NM_001008274) using the Wallace formula and Bolton and McCarthy formula with PCR product lengths in 150 ~ 300 bps, 500~800 bps and 800~1000 bps. a, accuracy (%); t, running time (ms). Boldface indicates highest values.

T_m formula and primer design methods	Wallace's formula				Bolton and McCarthy formula			
	PSO		PWLCPSO		PSO		PWLCPSO	
PCR product length	a (%)	t (ms)	a (%)	t (ms)	a (%)	t (ms)	a (%)	t (ms)
150~300 bps	64.6	370812	85.0	267375	45.0	463844	74.8	359250
500~800 bps	66.8	362219	89.8	234203	45.2	489078	70.0	375750
800~1000bps	52.0	454219	52.0	456078	29.2	544641	52.0	456078
average	61.1	395750	75.6	319218	39.8	488188	65.6	397026

7. REFERENCES

[1] BARANOVSKY, A. and DAEMS, D. 1995. Design of one-dimensional chaotic maps with prescribed statistical properties. *International Journal of Bifurcation and Chaos 5*, 1585-1598.

[2] CHEN, S.H., LIN, C.Y., CHO, C.S., LO, C.Z. and HSIUNG, C.A. 2003. Primer Design Assistant (PDA): A web-based primer design tool. *Nucleic Acids Res 31*, 3751-3754.

[3] HSIEH, M.H., HSU, W.C., CHIU, S.K. and TZENG, C.M. 2003. An efficient algorithm for minimal primer set selection. *Bioinformatics 19*, 285-286.

[4] HSIEH, S.T., SUN, T.Y., LIU, C.C. and TSAI, S.J. 2009. Efficient population utilization strategy for particle swarm optimizer. *IEEE Trans Syst Man Cybern B Cybern 39*, 444-456.

[5] K MPKE, T., KIENINGER, M. and MECKLENBURG, M. 2001. Efficient primer design algorithms. *Bioinformatics 17*, 214-225.

[6] KELLERT, S.H. 1993. *In the wake of chaos: Unpredictable order in dynamical systems*. University of Chicago Press.

[7] KENNEDY, J. and EBERHART, R. 1995. Particle swarm optimization. *IEEE International Conference on Neural Networks 4*, 1942-1948.

[8] LIANG, J.J., QIN, A.K., SUGANTHAN, P.N. and BASKAR, S. 2006. Comprehensive learning particle swarm optimizer for global optimization of multimodal functions. *IEEE Transactions on Evolutionary Computation 10*, 281-295.

[9] MIURA, F., UEMATSU, C., SAKAKI, Y. and ITO, T. 2005. A novel strategy to design highly specific PCR primers based on the stability and uniqueness of 3'-end subsequences. *Bioinformatics 21*, 4363-4370.

[10] MULLIS, K.B. and FALOONA, F.A. 1987. Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. *Methods Enzymol 155*, 335-350.

[11] SAMBROOK, J., FRITSCH, E.F. and MANIATIS, T. 1989. *Molecular cloning*. Cold Spring Harbor Laboratory Press Cold Spring Harbor, NY.

[12] SHI, Y., EBERHART, R.C., TEAM, E. and KOKOMO, I.N. 2001. Fuzzy adaptive particle swarm optimization. *Proceedings of IEEE International Conference on Evolutionary Computation 1*, 101-106.

[13] VAN DEN BERGH, F. and ENGELBRECHT, A.P. 2004. A cooperative approach to particle swarm optimization. *IEEE Transactions on Evolutionary Computation 8*, 225-239.

[14] WALLACE, R.B., SHAFFER, J., MURPHY, R.F., BONNER, J., HIROSE, T. and ITAKURA, K. 1979. Hybridization of synthetic oligodeoxyribonucleotides to phi chi 174 DNA: the effect of single base pair mismatch. *Nucleic Acids Res 6*, 3543-3557.

[15] WANG, J., LI, K.B. and SUNG, W.K. 2004. G-PRIMER: greedy algorithm for selecting minimal primer set. *Bioinformatics 20*, 2473-2475.

[16] WU, J.S., LEE, C., WU, C.C. and SHIUE, Y.L. 2004. Primer design using genetic algorithm. *Bioinformatics 20*, 1710-1717.

[17] XIANG, T., LIAO, X. and WONG, K. 2007. An improved particle swarm optimization algorithm combined with piecewise linear chaotic map. *Applied Mathematics and Computation 190*, 1637-1645.

[18] YANG, C.H., CHENG, Y.H., CHANG, H.W. and CHUANG, L.Y. 2010. Primer Design with Specific PCR Product using Particle Swarm Optimization. *International Journal of Chemical and Biomolecular Engineering 3*, 18-23.

[19] YANG, C.H., CHENG, Y.H., CHUANG, L.Y. and CHANG, H.W. 2009. Specific PCR product primer design using memetic algorithm. *Biotechnol Prog 25*, 745-753.

[20] YANG, X., YUAN, J., YUAN, J. and MAO, H. 2007. A modified particle swarm optimizer with dynamic adaptation. *Applied Mathematics and Computation 189*, 1205-121

Performance Analysis of Heterogeneous Computer Cluster

Abdusy Syarif
Information Engineering Study Program
Faculty of Computer Science
Mercu Buana University – Jakarta
abdusyarif@mercubuana.ac.id

Saiful Ikhwan
Information Engineering Study Program
Faculty of Computer Science
Mercu Buana University – Jakarta
thole_gdg@yahoo.com.sg

Muhammad Risky
Information Engineering Study Program
Faculty of Computer Science
Mercu Buana University – Jakarta
muhammad.risky@gmail.com

ABSTRACT
This paper measures the performance and load distribution among PC nodes in heterogeneous computer cluster system. The method used in this research is creating a program to count the value of Phi and modifying a module called MPI (Message Passing Interface). The computer cluster system tested in this research consists of five PCs on local area network. In conclusion, the computer cluster reduce processing time significantly for high iteration workload (10,000,000 iterations) but it is not effective for low iteration workload (5,000 iterations). For high workload, the performance of 5 nodes cluster is 3.8 times faster than 1 node cluster. However, for low workload, the performance of 5 nodes is 0.12 time slower than 1 node cluster.

Keywords
Computer Cluster, Workload, Time Consumption. The smallest error

1.INTRODUCTION
In the development of increasingly advanced computer technology, in terms of computer speed, speed of RAM, processor speed, an old computer could be useless in the future time. The latest modern computer purchased today could be an old stuff in a matter of months or years due to the emergence of new generation processor. Therefore, today's popular alternative is a cluster computer (computer group) or a parallel computer (parallel computers).

This computer cluster method is a group of computers connected to each other to work together in a local area network in order to solve a problem faster than a stand alone computer. Computer cluster sometimes called parallel computers also, because of its work in parallel to solve the problems simultaneously.

Researchers from ITS and Hiroshima University [6,7] have designed and implemented grid computing using Globus toolkits and Condor. In their research they used Message Passing Parameter (MPI) as communication protocol between nodes. And researchers from Dublin City University [8] performed clustering in solving problems in computer vision. In addition, they used ParallelKnoppix with MPI too.

2. COMPUTER CLUSTER
2.1 Parallel Processing
Parallel processing is the use of more than one CPU to run a program simultaneously. Ideally, parallel processing makes program run faster because more CPU is used.

Parallel computing is to perform computational calculation using 2 or more CPU / processor in a computer or in different computers, in this case each instruction is divided into a few instructions and then sent to all involved processors. The distribution of computing process is performed by a module called Message Parsing Interface (MPI).

Parallel processing includes:

- Information processing that underlines on elements data simultaneously.
- Intended to accelerate the computation of computer system and to increase the amount of output that can be produced in certain periods.
- Information processing that focused on manipulation of data elements owned by one or more process in order to solve a problem.

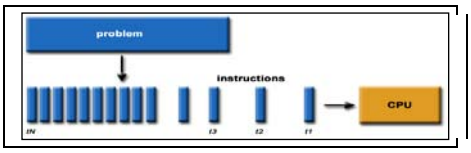


Figure 1. Single/Serial Computation [5]

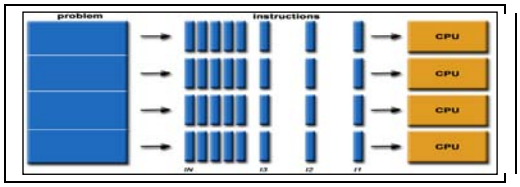


Figure 2. Parallel Computation [5]

2.2 Hardware
In building a computer cluster, firstly, we prepare the hardware to be used. The most important part of the computer cluster is computers and network devices. Because hardware is the core of computer cluster system, hardware should be able to work in a long time without stop or terminate. Hardware capabilities will be forced to work on maximum and stable.

Before we build a cluster computer, we should know the components. Main board is a place to put the components of other devices. Stability and compatibility main board is important, otherwise we will face trouble when installing components. Another significant hardware is processor. We use processors in accordance with the needs of the calculation or computing in cluster system. Other components that are considered to have great influence in the settlement computing is RAM or memory. Components of media data storage or hard disk is a place for the operating system and data to be stored. The next thing we need is a network card. This component is to connect and communicate between computers on the network. In addition we can use one monitor and one keyboard to control all the computers in cluster system through Local Area Network .

2.3 Heterogeneous

We built a computer cluster using Linux operating system and Message Passing Interface (MPI) with five computers interconnected by a switch with UTP straight cable. The specification of these computers or nodes are :

[1] Node1: Intel Pentium 4 1.7Ghz, 512MB RAM and 40GB Harddisk

[2] Node2: Intel Pentium 4 2.0Ghz, 512MB RAM and 40GB Harddisk

[3] Node3: Intel Pentium 4 2.0Ghz, 1GB RAM and 40GB Harddisk

[4] Node4: Intel Pentium 4 3.0Ghz, 512MB RAM and 40GB Harddisk

[5] Node3: Intel Pentium 4 3.0Ghz, 512MB RAM and 40GB Harddisk

Figure 3 below is computer cluster topology that we have built.

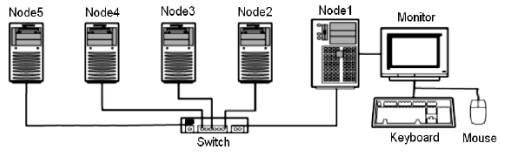


Figure 3. Network Topology cluster system

Figure 3 show Network topology in used because it is the most simple topology and easy installation. UTP cables used to connect all computers through a switch hub in local area network (LAN). IP Address for each computer can use a private Internet address. In this case , we provide IP addresses ranging from 192.168.1.1 for node1 and so on until 192.168.1.5 to node5. All nodes have subnet 255.255.255.0.

2.4 Software

The operating system that we use is Red Hat Linux 7.2 kernel 2.4.7-10. Furthermore, a module or software is needed, message passing interface (MPI). MPI is a parallel programming environment that very crucial in this case. After all hardware were installed properly, the next step is configuration. Actually, there are some software or operating systems to implement cluster computer, for example Globus, Condor, ParallelKnoppix, Amoeba, Angel, Chorus, GLUnix, Guide, Hurricane, Mach, Masix, Mosix, etc.

In this research, we used Red Hat operasting system and MPI with MPICH-1.2.5. For connection between nodes, we required a remote shell that can access all computer resources from the outside (remote) node. For this purpose we used rsh (remote shell) to communicate or configure nodes.

2.4.1 Message Passing Interface

Message Passing Parameter (MPI) a language-independent communication protocol used to program parallel computers. Both point-to-point and collective communication are supported. MPI "is a message-passing application programmer interface, together with protocol and semantic specifications for how its features must behave in any implementation."MPI's goals are high performance, scalability, and portability. MPI remains the dominant model used in high-performance computing today.

MPI is not sanctioned by any major standards body; nevertheless, it has become a de-facto standard for communication among processes that model a parallel program running on a distributed memory system. Actual distributed memory supercomputers such as computer clusters often run these programs. The principal MPI-1 model has no shared memory concept, and MPI-2 has only a limited distributed shared memory concept. Nonetheless, MPI programs are regularly run on shared memory computers. Designing programs around the MPI model (contrary to explicit shared memory models) has advantages over NUMA architectures since MPI encourages memory locality. Although MPI belongs in layers 5 and higher of the OSI Reference Model, implementations may cover most layers of the reference model, with socket and TCP being used in the transport layer. Most MPI implementations consist of a specific set of routines (i.e., an API) directly callable from Fortran, C and C++ and from any language capable of interfacing with such routine libraries (like C#, Java or Python). The advantages of MPI over older message passing libraries are portability (because MPI has been implemented for almost every distributed memory architecture) and speed (because each implementation is in principle optimized for the hardware on which it runs). MPI uses Language Independent Specifications (LIS) for the function calls and language bindings. The first MPI standard specified ANSI C and Fortran-77 language bindings together with the LIS. The draft of this standard was presented at Supercomputing 1994 (November 1994) and finalized soon thereafter. About 128 functions the MPI-1.2 standard in its present definition.

At present, the standard has got a couple of popular versions: version 1.2 (shortly called MPI-1), which emphasizes message passing and has a static runtime environment, and MPI-2.1 (MPI-2), which includes new features such as parallel I/O, dynamic process management and remote memory operations. MPI-2's LIS specifies over 500 functions and provides language bindings for ANSI C, ANSI Fortran (Fortran90), and ANSI C++. Interoperability of objects defined in MPI was also added to allow for easier mixed-language message passing programming. A side effect of MPI-2 standardization (completed in 1996) was clarification of the MPI-1 standard, creating the MPI-1.2 level.

2.4.2 Testing With Program 'Hitung'

A program 'hitung' was modified using parallel programming in order to prove whether all the cluster nodes can run the program simultaneously. The program used C language and compiled using `mpicc` command. The file result embedded in all nodes. This example program already exists in the package MPICH program. Because this system is a distributed computer system, so that must be the same file and in the same directory. By running the command "`~$mpirun -np 5 hitung`" then the output will appear as Figure 4.

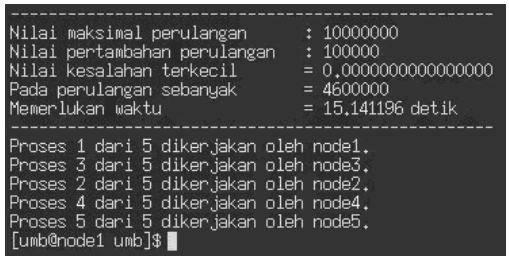


Figure 4. Result output program hitung for 5 processor

This program is a modified version of the MPI package. The Original one only count once process by displaying the value of error and processing time. In modification code of this program, this program loop the process and then compare the results of calculation to be taken with the smallest error values, the iteration of n times and processing time to complete the counting process. To test this program, we use a program to calculate Phi. This program will calculate the smallest error value by comparing several calculations of data obtained. This is the procedure.

```
double f(double a)
{
    return (4.0 / (1.0 + a*a));
}
h = 1.0 / (double) n;
sum = 0.0;
for (i = myid + 1; i <= n; i += numprocs)
{
x = h * ((double)i - 0.5);
sum += f(x);
}
mypi = h * sum;
```

2.4.3 Task Distribution

Task distribution is performed automatically by MPI. When there is a big task, then MPI will divided and distributed them into all nodes.

Table 1. Task Distribution 5 nodes

	node1	node2	node3	node4	node5
Task	1	2	3	4	5
	6	7	8	9	10
	11	12	13	14	15

Table 1 describes task distribution, task 1,6,11 performed by node1 and task 2,7,12 performed by node2 and so on.

3. RESULT AND DISCUSSION

With MPI, the processor's workload would be divided at each node.Table 2 is result of monitoring the processor load on node1 to node5 when program 'hitung' was run for iteration until 15,000,000 iterations.

Table 2. The results of processor workload monitoring.

Node Ke / Jumlah Node	1	2	3	4	5
1	90-94%	0-3%	0-3%	0-3%	0-3%
2	86-94%	81-89%	0-3%	0-3%	0-3%
3	78-90%	73-80%	87-98%	0-3%	0-3%
4	72-88%	69-85%	85-93%	82-89%	0-3%
5	68-84%	66-80%	76-89%	74-82%	80-88%

If node1 runs MPI program using only 1 node, then the processor load 90% till 94%. While other nodes do not join to run MPI program, load processor is between 0% to 3%. Node1 decrease processor load when running MPI program using 2 nodes with node2 is 86% to 94%. While on node2 reach 81% to 89%.

When running MPI program using the 3 nodes on node1, node2 and node3, node1 decrease in the load on the processor back to the 78% to 90%. On node2 while also decreasing the load on the processor that is 73% to 80% and 87% node3 reached up to 98%.

When running MPI programs using the 4 nodes on node1, node2, node3 and node4, node1 decrease in the load on the processor back to the 72% to 88%. On node2 while also decreasing load on the processor 69% to 85%. In node3 decrease the load on the processor 85% to 93% and 82% node4 reach up to 89%.

And when running MPI programs using all the nodes, the load decreased node1 processor that is at least 68% to 84%. On node2 also decrease the load on the processor that is 66% to 80% and the decrease node3 processor load is at 76% to 89%. In node4 decrease load on the processor 74% to 82% and 80% node5 reach up to 88%.

The smallest error we can get using 5 nodes on cluster computer shows in table 3 with iteration added value is 10 and iteration maximum is 5,000.

From the manual calculation results, we obtained Phi value by two iteration is 3.1510. In the same way we can do this action for n times to obtain the smallest error value. When using 2 or more nodes with more than one task, PC cluster will automatically divide tasks into another node.

Tabel 3. The Smallest Error and Time Consumption with min iteration 5,000.

Node	Iteration	Time	Smallest Error
1	5,000	3.487530 sec	0,0000000333332920

2	5,000	3.501618 sec	0,0000000333332690
3	5,000	3.508167 sec	0,0000000333332690
4	5,000	3.508390 sec	0,0000000033333336
5	5,000	3.918370 sec	0,0000000033333336

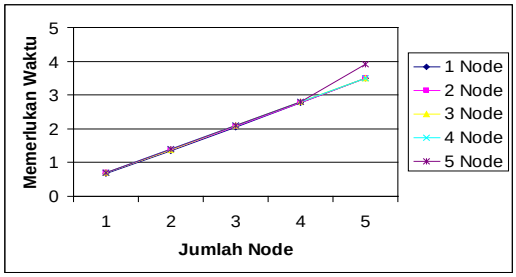


Figure 5.Time Consumption With Min. Iteration

When minimum iteration (5,000) was applied to program 'Hitung', both time consumption and smallest error are almost similar. However, when 5 nodes were taking part in cluster computer and trying to finish a small task, time consumption was higher than before. And it is cause the smallest error is not significant different. For instance, with 2 nodes in cluster system, with 5,000 iterations, it takes time 3.501618 seconds. And while 5 nodes, it takes time 3.918370 seconds. It means that with adding 3 nodes it only faster 0.416752 second or 11.9%.

Table 4. The Smallest Error With Max. Iteration 10,000,000

Node	Time	Iteration	Smallest Error
1	72.848994 sec	8,700,000	0,0000000000000018
2	36.792878 sec	5,700,000	0,0000000000000009
3	24.762437 sec	3,500,000	0
4	18.761777 sec	3,000,000	0,0000000000000027
5	15.141196 sec	4,600,000	0

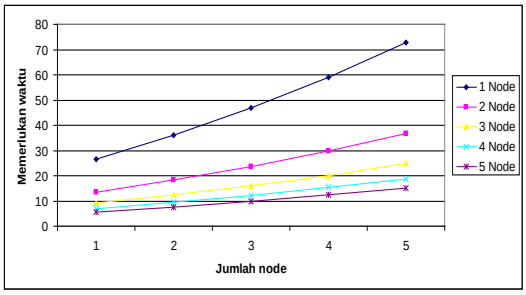


Figure 6. Time Consumption With Max. Iteration

Table 4 shows that there are significant different time consumption between 2 nodes and 5 nodes in cluster with maximum iteration (10,000,000). When 2 nodes in cluster, with 5,700,000 iteration, it takes time 36.792878 seconds, but when 5 nodes in cluster, it takes time only 15.141196 seconds. It means reduce time consumption until 21.651682 seconds or about 58.84%.

4. CONCLUSION

In cluster computer system when there is a process, then the process will be directly distributed to all nodes dependent on (command: ~$mpirun-np n mpi) option used, for example 2, 3, 4 or 5 nodes. CPU load changes do not occur when node2 clustering process 0% to 3% to 81% to 89% when plated on the 2 nodes. The load CPU of each CPU (5 nodes) up to 68% to 84%. This is because the tasks are divided into the CPU which is connected in a computer cluster, while if the PC to work independently CPU load reaches 100%. CPU Load smallest occurs when using a 5 node, and the largest CPU load occurs when using 1 node.

When program 'Hitung' performed in small or minimum iteration, more node do not decrease time consumption. As consequently, the smallest error do not significant different from the others. It caused by delay on network process. However, when in high or maximum iteration, more node could decrease time consumption.

5. REFERENCES

[1] Hiraki Laboratory, 2008, *OS Projects* [online], available http://www-hiraki.is.s.u-tokyo.ac.jp/members/nobukunifull.html [Accessed 8 November 2009]

[2] Red Hat Documentation , 2001, The Official Red Hat Linux x86 Installation Guide [online], available: http://www.redhat.com/docs/manuals/linux/RHL-7.2-Manual/install-guide/ [Accessed 8 November 2009]

[3] Amar, 2008, Boot Loader [online], available: http://amaronly.blogsome.com/2008/01/15/bootloader/ [accessed 8 November 2009]

[4] Supriyadi, Andi, 2007,"Choosing Network Topology and Hardware in Designing Computer Network", http://www.litbang.deptan.go.id/warta-ip/pdffile/4.andidhani_ipvo l16-2-2007.pdf [accessed 8 November 2009]

[5] Blaise Barney, “Introduction to Parallel Computing”, https://computing.llnl.gov/tutorials/parallel_comp/ , 13 September 2007

[6] M. Hariadi, Arief Kurniawan, Nur Kholis M, Tohru Kondo,”*Design and Implementation Grid Computing Environment Using Globus Toolkit*”, Seminar on Intelligent Technology and Its Application, ITS, Surabaya, 2008

[7] Arief Kurniawan, M. Hariadi, Lukmanul Hakim, Tohru kondo,”*Design and Implementation of High Throughput Computing Environtment Using Condor*”, Seminar on Intelligent Technology and Its Application, ITS, Surabaya, 2008

[8] Ade Jamal, Pandriya Sistha,”*Performance Data Communication Broadcast Collective on PC Cluster*”, Seminar on Computing in Science and Nuclear Technology XVII, Batan, 2006

[9] Andrew Fiade,”*Performace Analysis of Cluster Between Globus and Alchemi*”,Thesis, Faculty of Computer Science, University of Indonesia, 2008

Reduced Space Classification Using Kernel Dimensionality Reduction for Question Classification in Public Health Question-Answering

Hapnes Toba
Maranatha Christian University
Bandung, Indonesia / University of Indonesia Depok, Indonesia
hapnes.toba@eng.maranatha.edu / hapnes.toba@ui.ac.id

Ito Wasito
Information Retrieval Laboratory
University of Indonesia
Depok, Indonesia
ito.wasito@cs.ui.ac.id

ABSTRACT
One of the major problems in Question Answering System is how to classify a question into a particular class that further will be used to find exact answers within a large collection of documents. Kernel Dimensionality Reduction (KDR) is an alternative method that can be used for features reduction, and in the same time classify question type by using the most effective m-dimensional features in its vector space. In this experiment we used question-answer pairs data from public health domain and word (unigram) features construction. This research shows that KDR correct rate performance is better than SVM after a head-to-head comparison from 100 observations.

Keywords
Kernel Dimensionality Reduction, Reproducing Kernel Hilbert Space, Supervised Machine Learning, Question Classification, Question Answering System

1. INTRODUCTION
Question answering system (QAS) is a form of information retrieval that used a natural language question as its input and returns explicit answers in the form of a single answer or snippets of text rather than a whole document or set of documents. One of the most challenges in QAS is how to classify a question into a particular class that further will be used to find exact answers within a large collection of documents. Two major approaches has been widely used in question classification, i.e. the pattern-based and machine learning approach [1]. While the pattern-based approach try to identify a question in its syntax form which can be resource intensive [2], on the other hand machine learning approach try to approximate in which class a question can be classified by using an already trained classifier [3], [4], [5].

The widely used algorithm in question classification using the machine learning approach are mainly based on supervised classification using the Support Vector Machines (SVM) [6], [7], [9], and Maximum Entropy [3], [4], [8] to analyze the semantic and syntactic structure. Due to the data sparsity and features selection problem in both algorithms, it is hard to choose the best features in a particular question class. In this paper we will introduce that Kernel Dimensionality Reduction (KDR) algorithm can be used to reduce word matrix features and in the same time classify question type using the most effective m-dimensional features in its vector space. The rest of this paper is organized as follow: section 2 will give an exploration of KDR and the feature selection method. Our research design will be described in section 3, followed by the experiments and their results in section 4. Some discussions, conclusions and future works will be presented in section 5.

2. METHODS EXPLORATIONS

2.1 KDR with Reproducing Kernel Hilbert Spaces
KDR is based on a particular class of operators on reproducing kernel Hilbert spaces (RKHS) [10]. A Hilbert space is an extension of a vector space. It requires the definition of an inner product on the vector space [15] which enables it to be called an inner product space. An example of an inner product on a finite vector space between any vector $\underline{x}$ and $\underline{y}$ is:

$$(x, y) = \sum_{i=1}^{n} xi.yi$$

The KDR algorithm relates dimensionality reduction to conditional independence of variables, and use RKHS to provide characterizations of conditional independence and thereby design objective functions for optimization. The hypothesis is to find effective subspace that can be formulated in terms of conditional independence. In particular, it is assumed that there is an r-dimensional subspace $S \subset R^m$ such that the following equality holds for all x and y:

$$pY|X(y|x) = pY|\prod sX(y|\prod sx) \quad \ldots (1)$$

Let (A,B) be an m-dimensional orthogonal matrix such that the column vectors of A span the subspace S (so A is m × r), and define $U=A^TX$ and $V=B^TX$. Because (A,B) is an orthogonal matrix, we can derive that $p(X(x))=p(U,V(u,v))$ and $p(X,Y(x,y))=p(U,V,Y(u,v,y))$.

Eq. (1) is thus equivalent to:

$$pY|U,V(y|u,v) = pY|U(y|u) \quad \ldots (2)$$

In this way, the effective subspace S is the one which makes Y and V conditionally independent given U [10] (see Figure 1).

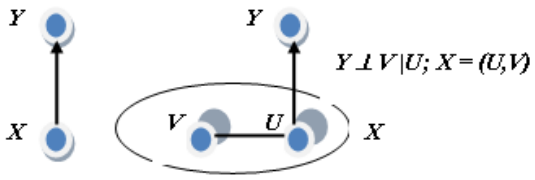


Figure 1. Representation of Dimensionality Reduction [10]

Another important viewpoint on the equivalence between conditional independence and the effective subspace is the mutual information condition that holds the dimensionality reduction. It is known that:

$$I(Y,X) = I(Y,U) + E_u[I(Y|U,V|U)] \quad \ldots (3)$$

where `I(Y,X)` is the mutual information between X and Y. Because Eq. (1) and (2) implies `I(Y,X)=I(Y,U)`, the effective subspace S is characterized as the subspace which retains the entire mutual information between X and Y, or equivalently, such that `I(Y|U,V|U)=0`. This produces again the conditional independence of Y and V given U.

KDR uses covariance operator on RKHS to produce an objective function for dimensional reduction. If there is a set Ω consisting of feature vectors in its columns, RKHS is produced by using the kernel that has the following reducing property: `<f,i(.,x)>H=f(x)` for all `x` the elements in the vector space and all `f` the functions (or features in this sense) in `H`, the reproduced space. Fukumizu et. al. in [10] uses the Gaussian kernel $i(x1,x2)=\exp(-||x1-x2||^2/2\sigma^2)$. We will have thus `(H,i)` which is a reproducing kernel Hilbert space of functions on a set of random vectors in Ω with a positive definite kernel $i:\Omega\times\Omega\rightarrow R$ and an inner product `<.,.>` in H. The vector space that has been reproduced by the kernel function need to be further processed to guarantee the conditional probability and linear independency of the reduced kernel. This is achieved in KDR by using the cross covariance operator ΣYX from H1 to H2 that defined by the relation:

$$\langle g, \Sigma_{YX} f \rangle_{\mathcal{H}_2} := E_{XY}[f(X)g(Y)] - E_X[f(X)]E_Y[g(Y)] \quad (= \mathrm{Cov}[f(X), g(Y)]) \quad \ldots (4)$$

This relation implies that the covariance of `f(X)` and `g(Y)` is given by the action of the linear operator ΣYX and the inner product. Interested readers should refer to [10] for the complete mathematical proofs.

2.2 Feature Selection

The features in the m-dimensional space of documents are usually formed by its textual features. Information retrieval research suggests that word stems can be used effectively as representation units of a document. Such word stems are derived from the occurrence form of word by removing case and flections information [11]. This leads to an attribute-value representation of text. Each distinct word w_i (unigram feature) corresponds to a feature with term frequency $TF(w_i,x)$, the number of times word w_i occurs in the document x, as its value.

Refining this basic representation, it is better to scale down the dimension of feature vector with their inverse document frequency $IDF(w_i)$ [12], which can be calculated from the document frequency $DF(w_i)$ which is the number of documents the word w_i occurs in:

$$IDF(w_i) = log(\frac{n}{DF(w_i)}) \quad \ldots (5)$$

Where n is the total number of documents. In this research we assume that a question is comparable as a document, and thus we called our feature as inverse question frequency of w_i, $IQF(w_i)$.

Our word matrix representation will have i-rows, that equal the number of questions and j columns, which equal the number of features (see Figure 2). The problem with such representation is the sparsity of data in each feature (j-th column) that represent the occurrences of a term in the all i-questions. It is reasonable that not every word should be appeared in all questions. This kind of problem will be useful to evaluate the performance of KDR and compare it with other comparable supervised method, in this case the support vector machines.

	IQF-1st	...	IQF-jth
row 1-st			
.			
.			
.			
row i-th			

Figure 2. Matrix Representation of Questions

2.3 Support Vector Machines

Support vector machines (SVM) are based on the Structural Risk Minimization principle from computational learning theory [13]. Joachims in [14] described the idea of SVM as structural risk minimization that try to find a hypothesis h for which we can guarantee the lowest true error. The true error of h is the probability that h will make an error on an unseen and randomly selected test example. An upper bound can be used to connect the true error of a hypothesis h with the error of h on the training set and the complexity of H (measured by VC-Dimension), the hypothesis space containing h. Support vector machines find the hypothesis h which (approximately) minimizes this bound on the true error by effectively and efficiently controlling the VC-Dimension of H. The SVM will thus in particular define the criterion to be looking for a decision surface that is maximally far away from any data point [16]. This distance from the decision surface to the closest data point determines the margin of the classifier. This method of construction necessarily means that the decision function for an SVM is fully specified by a subset of the data which defines the position of the separator. These points are referred to as the support vectors [16, 17] (see Figure 3).

Both KDR and SVM are promising to be compared because each method can handle the text classification properties, i.e.: high dimensional input space, few irrelevant features, document vectors are sparse, and most text categorization problems are linearly separable [14].

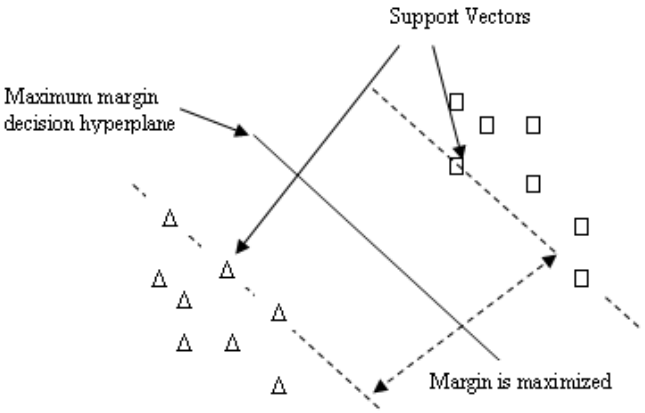


Figure 3. Maximization of margin of the support vectors [16]

3. EXPERIMENTAL SETTING

3.1 Algorithms and Tools

During the experiments the following algorithms and tools are used:

1. KDR algorithm (© Fukumizu 2004);
2. SVM Linear and RBF Classification (© 2004-2007 The MathWorks, Inc) ;
3. KNN Classification (© 2004-2007 The MathWorks, Inc).

3.2 Data

We used questions from the public health domain. We downloaded the question-answer pairs from the Singapore Ministry of Health FAQ pages[1] in the topics of swine flu (H1N1-2009) and gastric flu. In total there are 92 questions from those two topics (73 questions about H1N1 and 19 questions about gastric flu). The reason we have chosen those topics based on the assumption, that topics that share the same context (in this sense the "flu" context), will share the same features. This assumption is important to form an objective orientation when we try to classify a test question in classifier that was constructed from the same (randomize) dataset.

After we downloaded the FAQ, to obtain the version in Bahasa Indonesia, we used the Google translation tools[2]. The translation that we obtained was not directly used for the research. We reconstructed first some of the grammar and unmatched contextual terms that is used in daily Indonesian. After we have the final version of the translated FAQ in Bahasa Indonesia, we use Perl programming language to convert the FAQ into the feature matrix as described in section 2.2. We have thus for the feature matrix 92 rows and 1137 columns as features.

3.3 Performance Evaluation

In the benchmarking step, we trained first a classifier using the Linear and RBF SVM classification, and then tested it with random test data from a subset of the row data.

The complete procedures of the benchmarking steps are:

1. Using the whole matrix representation, we trained the data with SVM to separate two distinct classes using the [train, test] composition of [90, 10].
2. We tested the SVM classifier with the random generated test data from number 1, and find the correct rate for each composition in 100 runs.
3. We use the correct rate to evaluate the performance of each classifier, as follow:

$$CorrectRate = \frac{NumberOfCorrectClassifiedQuestions}{NumberOfTotalTestQuestions} \quad \ldots (6)$$

4. We saved the test-indexed question which will be used as the supervision vector in the KDR method.
5. We reduced the features matrix representation using KDR using the 2-dimensional reduction, 100 iterations and 0.1 learning rate.
6. Use the result of the already reduced KDR matrix as the input vector for the SVM training. In this step we use the concept of "build classifier in the reduced space", as also described in [10].
7. We use the saved test-indexed question (number 4) as the same test data for KDR classification.
8. Compare the results of the original SVM and the enhanced KDR results.

In our research, besides comparing the whole matrix, we also compare the KDR with "manual"-reduced SVM. This "manual"-reduced SVM, is a reduced matrix that was formed by selecting the two most occurrence words that occur in all question classes after we applied the stemming and removed the stop words, and used them as the features. To compare the resulting KDR classification, we also took the KNN classifier [18], with K=2 and K=5, to see how close the distance among the classified question-answer pairs.

To compare the consistency of the KDR and the SVM classification, we used the head-to-head comparison over 100 evaluation runs on both methods, i.e. we run the experiment 100 times for each method and then count how many times a method outperforms the other. We also compute the mean and standard deviation for each method to see the performance in overall evaluation runs.

4. EXPERIMENTS AND ANALYSIS

We have run our experiments according to the setting which is described in Section 3.

4.1 KDR Iterations and Dimensions

The purpose of this experiment is to see the impact of KDR number of iterations and dimension construction. We run an experiment that used 50, 100 and 500 iterations to construct a dimension of 2, 5, 10, 15, 20, 50, 100, 125, 150 and 500. The result of this experiment can be seen in Figure 4.

[1] http://www.pqms.moh.gov.sg/apps/fcd_faqmain.aspx (menu: Illness and Diseases), accessed on February 2010

[2] http://www.google.co.id/language_tools?hl=id

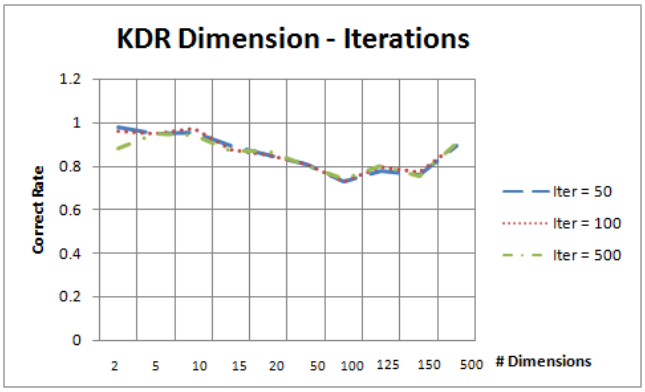


Figure 4 Performances of KDR Iterations and Dimensions

From Figure 4, we can see that the performances of 2- and 10-dimension reduced matrix are the best for all iterations. This indicates that KDR can still perform it best in a small number of dimension (features) which is important in the benchmarking steps (cf. section 3.3). We can also see from Figure 4, that the correct rate patterns are almost identical for each iteration. This result indicates that the number of iterations has no direct impact to the number of dimensions. Figure 5 plots the impact of number of iterations (Blue = H1N1, Yellow = gastric).

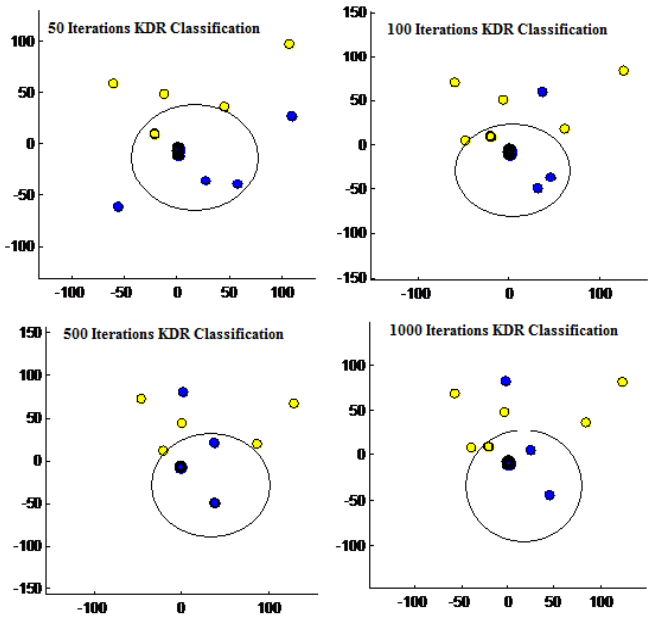


Figure 5. Vector Distribution of KDR Classification (50, 100, 500 and 1000)

The number of iterations indicates how fast the learning rate closer to a convergence area in each training-classification session. Based on the results in this experiment, we choose the 2 dimensions and 100 iterations as our default setting in the benchmarking steps.

4.2 SVM and KDR Classification

We used the SVM classification with linear and RBF function (sigma = 1). The resulting "mean value" correct rate classification of each random generated [90, 10] composition for all 1137 features in four series of 100 training-classification runs can be seen in Table 1. For the "manual" reduced SVM, the most occurrence words that occur in both classes H1N1 and gastric flu are the word "*flu*" and "*virus*".

Table 1 SVM Results for All Features

SVM all features		SVM Most 2-Words (flu-virus)	
RBF Sig=1	Linear	RBF Sig=1	Linear
0.8933	0.8867	0.8967	0.8822
0.8911	0.8744	0.8844	0.8933
0.8811	0.8978	0.8922	0.8689
0.8900	0.8878	0.8967	0.8856

Result in Table 1 shows that "all features" SVM Linear classification performed better than when we choose only "several selected features". To compare the performance of Linear and RBF SVM from Table 1 against the KDR algorithm, we run KDR experiment with 2-dimensional features in 50, 100, and 500 iterations in four series of 100 training-classification runs, which further classified using the RBF and linear SVM classifier. The "mean value" of the correct rate in these experiments can be seen in Table 2. Results of these KDR experiments show again that the iterations number does not give any direct impact to the correct rate (cf. section 4.1).

Table 2. KDR Performance

KDR (2-dim, 50 iter)		KDR (2-dim, 100 iter)		KDR (2-dim, 500 iter)	
RBF Sig=1	Linear	RBF Sig=1	Linear	RBF Sig=1	Linear
0.9783	0.9783	0.9722	0.9783	0.8261	0.9783
0.9722	0.9783	0.9783	0.9722	0.8056	0.9722
0.8701	0.9565	0.8701	0.8837	0.9565	0.913
0.8837	0.8701	0.9588	0.9565	0.8701	0.8837

Interpretation plot from the result in Table 1 and 2 can be seen in Figure 6a and 6b.

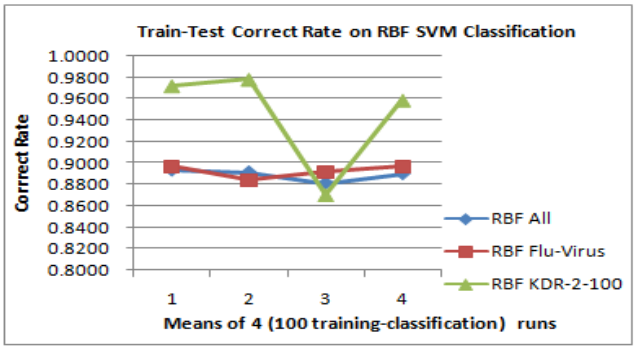


Figure 6a. Comparison of Correct Rate RBF Linear SVM (right) on Different Train-Test Composition

Those figures give us an insight that KDR reduction matrix which is trained using the RBF-SVM and linear-SVM has given an almost identical correct rate patterns during the 4 series of training-classification runs. Such result is also hold for the original "all features" and the "manual" constructed features.

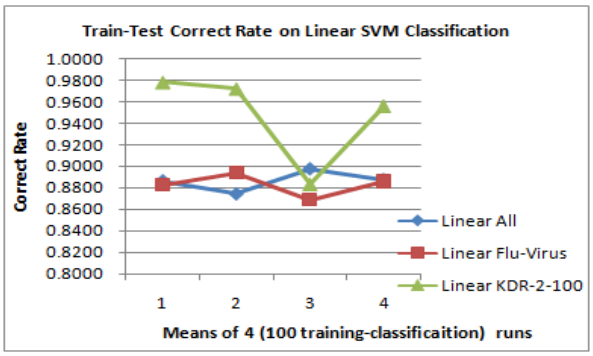


Figure 6b. Linear SVM (right) on Different Train-Test Composition

4.3 Effectiveness of KDR

To see how effective the dimension reduction in KDR, we also plotted the "manual" constructed 2-dimension features and the KDR 2-dimension. The plot can be seen in Figure 7a and 7b.

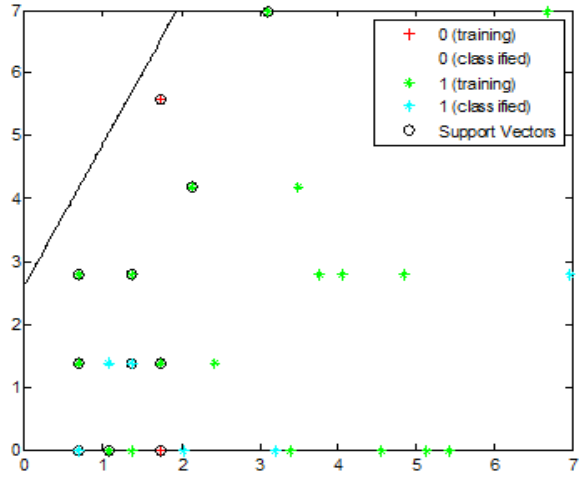


Figure 7a. 2-"Manual" Selected Features Linear SVM

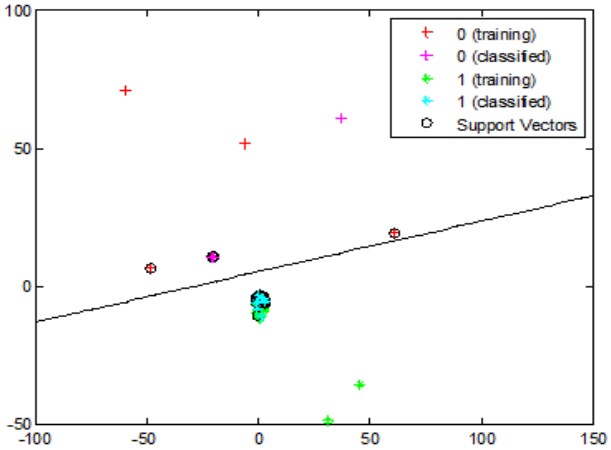


Figure 7b. KDR 2-dim 100- iterations Classification

Figure 7a & 7b give an indication that KDR is very effective to classify a huge number of features into a much smaller dimensions. KDR (Figure 7b) has produced better classification than the "manual" selected features (flu-virus in Figure 7a).

4.4 KDR and KNN Classification

The comparison of KDR and KNN gives another view of the KDR classification. Besides the reduced dimension that has been achieved with KDR, it also produces the classification that comparable with KNN. We used the data from the "manual" reduced features for our 2-NN and 5-NN classification. Figure 8 shows the plot of our experiment with 5-NN Euclidean Distance Classification compare with KDR (100 iterations). The x and y-axis the vector value estimations of the classifications.

Figure 8 shows us that the distance between KDR classified vectors is much closer than the KNN classification. In other words, this means that KDR can classify the features in the right classification although the distances between the features are very close one to another.

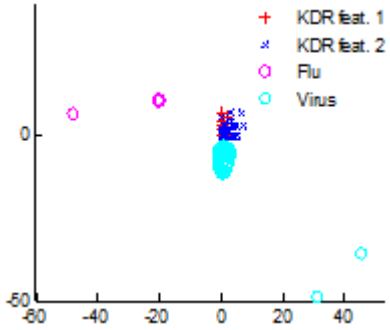


Figure 8. Comparison of Euclidean Distance (K=5) with "manual" selected features (flu-virus) and KDR (100 iterations)

4.5 Overall Performance Evaluation

To observe the consistency of KDR, we evaluated the correct rate performance of SVM Linear and RBF with all features, SVM Linear with manual constructed features (virus-flu), against the KDR-2-dim-100 iterations, in 100 training-classification runs with 10% holdout test data. The overall performance of the evaluation runs can be seen in Table 3. The graphical interpretation of each method in this evaluation can be seen in Figure 9.

Table 3. Overall Performances in 100 Evaluation Runs

	SVM Lin	SVM RBF	SVM Man	KDR
Mean	0.882222	0.878889	0.813333	0.958889
Std Dev	0.080248	0.080674	0.073872	0.060457
Mean +	0.96247	0.959563	0.887205	1.019346
Mean -	0.801974	0.798215	0.739462	0.898432

Because the value of mean +/- the standard deviation of each method is overlapping (see also Figure 9), the result is not conclusive. We need thus to evaluate the head-to-head comparison. This comparison gives an insight about the performance of each method in each evaluation run. We will see how many times a method outperforms the other.

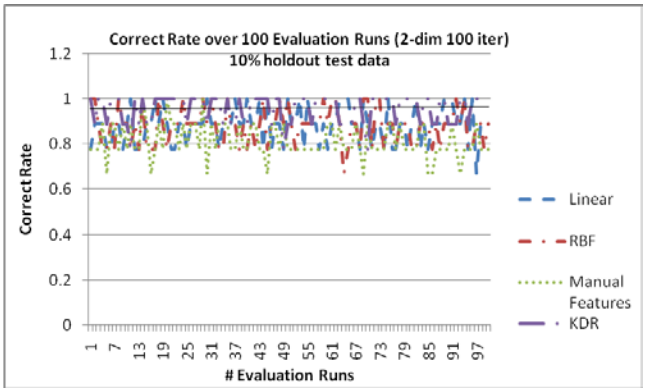


Figure 9. Correct Rate over 100 Evaluation Runs

The result of the head-to-head comparison of 100 observations can be seen in Table 4.

Table 4. Head-to-Head Comparison

Comparison	Lin	RBF	Man	KDR
Lin	0	66	86	37
RBF	34	0	90	39
Man	14	10	0	13
KDR	63	61	87	0

Each row in Table 4 gives the number of "winning" or equal correct rate of each method against the other. We can see from Table 4 that KDR classification outperforms the other methods. The "manual" constructed features perform the worst in each method; it indicates that such subjective selected features should be strengthened with some other features which will give better classification.

5. CONCLUSIONS & FUTURE WORKS

We found that KDR can be used as a promising alternative method to classify questions in Question Answering System. An important viewpoint is that KDR can effectively classify questions even with only very view features (words), i.e. 2-dimensions (cf. Section 4.1). KDR can also determine the best effective features in the vector space. The classifications of questions using the features reduction that KDR has determined in most of the time are better than the "manually" constructed features and the original "all feature" matrix (cf. Section 4.2 and 4.3). During the head-to-head comparison, we found that KDR outperforms significantly the SVM classification in many cases. This indicates that the features reduction that has been produced by KDR is very effective to be used in classification of questions (cf. Section 4.4 and 4.5).

As future works, we are going to apply KDR to strengthen the question classification and answer validation method in our ongoing research to build an Indonesian Question Answering System. In this sense, we are going to build a KDR classifier that can be used to anticipate a set of important features (words) from a question which could be classified into more than one question class (multi-labeling). The classification that produced by KDR will be important to find the real context of the question.

6. REFERENCES

[1] Purwarianti, et. al. 2006. Estimation of Question Types for Indonesian Question Sentence. Department of Information and Computer Sciences, Toyohashi University of Technology.

[2] Toba, H & Adriani, M. 2009. Pattern Based Indonesian Question Answering System. Proceedings of the International Conference on Advanced Computer Systems and Information Systems (ICACSIS) University of Indonesia.

[3] Ittycheriah, A. et. al. 2001. IBM‟s Statistical Question Answering System. Proceedings of the 10th Text Retrieval Conference (TREC 2001).

[4] Schlaefer, Nico. 2007. Deploying Semantic Resources for Open Domain Question Answering. Diploma Thesis. Language Technologies Institute School of Computer Science Carnegie Mellon University.

[5] Li, X., Roth, D. 2002. Learning Question Classifiers. Proceedings of the 19th International Conference on Computational Linguistics, Taipei, Taiwan.

[6] Cruchet, S, et. al. 2008. Supervised Approach to Recognize Question Type in a QA System for Health. MIE.

[7] Joachims, T. 1999. Transductive Inference for Text Classification using Support Vector Machines. International Conference on Machine Learning (ICML).

[8] Berger, A. et. al. 1996. A Maximum Entropy Approach to Natural Language Processing. Computational Linguistics 22(1).

[9] Zhang, D & Lee, WS. 2003. Question Classification using Support Vector Machine. ACM SIGIR 2003.

[10] Fukumizu, K., Bach, F.R., Jordan, M.I. 2004. Kernel Dimensionality Reduction for Supervised Learning with Reproducing Kernel Hilbert Spaces. The Journal of Machine Learning Research. Volume 5. December 2004. Pages 73-99.

[11] Porter, M. 1980. An Algorithm for Suffix Striping. Program (Automated Library and Information Systems). Volume 14. Number 3. Pages 130-137.

[12] Salton, G. & Buckley, C. 1988. Term Weighting Approaches in Automatic Text Retrieval. Information Processing and Management. Volume 24. Number 5. Pages 513-523.

[13] Vapnik, V.N. 1995. The Nature of Statistical Learning Theory. Springer, New York.

[14] Joachims, T. 1998. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. Proceedings of the European Conference on Machine Learning, Springer.

[15] Rijsbergen, van C.J. 2004. The Geometry of Information Retrieval. Cambridge University Press.

[16] Manning, C.D., et. al. 2008. Introduction to Information Retrieval. Cambridge University Press.

[17] Cristianini, N., and Shawe-Taylor, J. 2000. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, First Edition (Cambridge: Cambridge University Press). http://www.support-vector.net/.

[18] Mitchell, T. 1997. Machine Learning. McGraw Hill.

The Developing of Interactive Software for Supporting the Kinematics Study on Linear Motion and Swing Pendulum

Liliana
Universitas Kristen Petra
Siwalankerto 121
Surabaya
(62) 031 2983000
lilian@petra.ac.id

Kartika Gunadi
Universitas Kristen Petra
Siwalankerto 121
Surabaya
(62) 031 2983000
kgunadi@petra.ac.id

Yonathan Rindayanto Ongko
Universitas Kristen Petra
Siwalankerto 121
Surabaya
(62) 031 2983000

ABSTRACT

Many students assume that physics which is learned at school is considered as a difficult and boring subject. This condition happens because of conventional teaching and lack of visual aids. As a result, the subject is considered boring, not attractive, and difficult to understand. This software is developed to make the student to be more enthusiastic in dealing with physics, not only in schools but also in their personal computer.

This software can help the student in order to learn about kinematics of rectilinear motion and swing pendulum. This software is developed using Borland Delphi 7.0 as interface, Macromedia Flash MX as material delivery, and Microsoft Access as database. This database is used to save the question in the quiz to make it possible to be changed.

Based on the polling result, it is concluded that 80% of the senior high school students 10^{th} and 11^{th} grade agreed that materials in this software is more attractive from the aspect of text layout, picture, and supporting simulation. Besides, 90% of senior high school students said that this software could increase the motivation in learning physics.

Keywords

Computer Aided Learning, Interactive Software, Physics, Rectilinear Motion, Pendulum Swing.

1. BACKGROUND

In our daily life, we do motion, such as walk, run or we do motion with tool, such as drive a car. Some motions have certain characteristics. For example, a stright forward with a constant speed. Although very familiar and simple, high school students have some difficulty to understand motion because it can't be imagine easily. To cover this problem, this research designs a tool to help the high school student learning motion by simulating the motion.

This tool is a *Computer Aided Learning* (CAL) software. Commonly, CAL supports students tutorial on a certain topic, so they can learn by themselves. Since the CAL wants to provide a complement material for students, CAL needs something doesn't teached at school. To learn physics well, student want to do some experiments. In this way, CAL will help students to see motion, instead imagine it.

In our developed software, we mean it as an interactive learning method. This software will allow students read tutorial about motion, do an evaluation test, practice motion calculation based on newton laws, and simulate the motion based on some given input.

This kind of learning method has been developed in education recent decades. Some survey on the advantages using this method, said that CAL will increase student's interest on the topic[1]. Usage tools such as CAL also makes the education more effective and efficient than conventional education. It because Cal not only gives student material on the topic of interest, but also show picture and simulation which can explain more detail than words.

Actually, CAL is developed based on human's study method. Some students would like study by reading books while the others by seeing and doing something[2]. Some students will satisfy just with knowing while the others will satisfy with doing. CAL is a tool which help students learning by knowing and doing so they will understand the topic wholely.

2. LINEAR MOTION

Linear motion is motion on a stright path[3],[4]. There are two major kinds of linear motion, uniform linear motion and non-uniform linear motion. Uniform linear motion has constant velocity. Non-uniform linear motion has constant acceleration. The total distant the motion gains can be calculate using eq 1 while the one of non-uniform linear motion can be calculate using eq 2.

$$s = s0 + v.t \quad (1)$$

Where s is the distant reached by the moving object and s0 is the original position before the object moves. v is velocity, t is time needed by the moving object to finish the distant.

$$s = v0.t + \frac{1}{2}at^2 \quad (2)$$

Where v0 is velocity before the object moves with acceleration. The acceleration itself can be calculated using the equation below.

$$a = \frac{vt - v0}{t - t0} \quad (3)$$

Where vt is velocity at observed time t while t0 is starting time when the non linear moving begins. Actually, non linear moving is horizontal moving. This kind of moving can be expanded become vertical moving, such as something falling from above and parabolic moving, such as throwing something above, and after that thing reaches the peak position, it will fall down. The expanding moving can be calculated using the generic equation below.

$$Sx = 2.\frac{v_o^2}{2g}\sin 2\alpha \qquad (4)$$

$$Sy = \frac{v_o^2 \sin 2\alpha}{g} \qquad (5)$$

Where sx and sy are the distant at horizontal and vertical direction respectively. g is gravity and α is the degree if the object moves in parabolic direction.

3. SWING MOTION

3.1 Single Pendulum

Swing motion is a motion of an object with distinct mass to left and right direction[5],[6]. For example, a ball swing using a thread as shown in Figure 1.

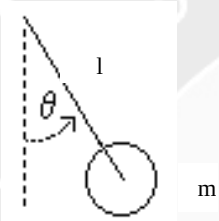


Figure 1. Single pendulum.

ln Figure 1, θ represents angle between pendulum with vertical line in degree. l is thread's length, in meter. To calculate the pendulum's position while swing, is used vector i and j as vertical and horizontal direction representation[5]. The pendulum's position can be calculated using equation 6. While period and frequency of the swing can be calculated using equation 7 and 8 respectively.

$$\text{position} = l \sin\theta\,\mathbf{i} - l\cos\theta\,\mathbf{j} \qquad (6)$$

$$\text{period} = \frac{2\pi}{\sqrt{g/l}} \qquad (7)$$

$$\text{frequency} = \frac{1}{2\pi}\sqrt{g/l} \qquad (8)$$

where g is earth's gravity. The Force of swing pendulum will be calculated using equation 9.

$$F = m.g.\sin\theta \qquad (9)$$

3.2 Double Pendulum

Double pendulum is two pendulum in one thread. They are tied sequently, as shown in Figure 2[6]. Double pendulum's motion represents one of chaotic motion in simple physics. Double pendulum in this CAL is focused on motion with undumping factor. It means motion without any external force works on it.

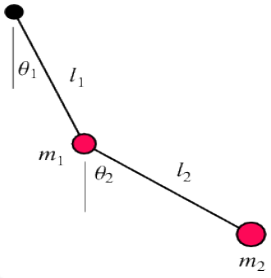


Figure 2. Double pendulum

To detect the position of first and second pendulum, used equation 10 and 11. Each position will be represent using vector. x1 and y1 is position of first pendulum, x2 and y2 is position of the second pendulum. l1 is thread's lenght from base to first pendulum while l2 is thread's length from first to second pendulum.

$$x_1 = l_1 \sin\theta_1$$

$$y_1 = -l_1 \cos\theta_1 \qquad (10)$$

$$x_2 = l_1 \sin\theta_1 + l_2 \sin\theta_2$$

$$y_2 = -l_1 \cos\theta_1 - l_2 \cos\theta_2 \qquad (11)$$

4. ANALYSIS

This CAL is designed for students to learn by themselves. They can make their own schedule, when they will finish their tutorial, they can test their ability by doing the evaluation test by theirselves. If they failed in mastering a distinct part, they can re-learn the part without annoying anyone else. None will wait other student's understanding in the studying process.

Firstly, this CAL wants to accomodate all the student learning behaviours. They will learn not only by reading the material but also hearing the explanation and seeing some support pictures. In this CAL, we adopt S-O-R method. In this method, studying process is started with receiving information, manipulating that information and finally, producing the result[7]. Information is given in text and voice form. Information is arranged based on what-why-how concept[8]. Firstly, give the definition, then, inform the background about the topic, explain equations which are related with, and finally, show how the equations work. After receive the information, student will be guided to manipulate the information. Through showing some examples and their calculations, and also simulate some equations, student will be supported to manipulate information they received before. The last step of S-O-R method is producing the result. This step is done by giving them some test to evaluate their understanding on the topic.

5. EXPERIMENTS

We had done some experiments to test the simulations. Figure 3 shows the interface of linear motion simulation menu. First user must input the velocity. Then the software will run a ball rolling on a linear line. While it rolls, we will see the distant it reaches each second. In this experiment, we input velocity 5 m/s. This simulation is done for five times. It reaches a distinct distant. Then, the time it needs to reach the distant is calculated using stopwatch, the result are:

Experiment 1 = 21.08 s

Experiment 2 = 21.14 s

Experiment 3 = 20.23 s

Experiment 4 = 20.10 s

Experiment 5 = 20.78 s

The average time is 20.65 s, while according to the equation, it should be 20 s. The error is less than 5%.

Figure 3. The interface of linear motion simulation menu.

To test non-uniform linear motion, we set time with 25 s, acceleration 2 m/s^2 and velocity 10 m/s. This simulation can be seen in Figure 4.

Figure 5 shows the result of simulation on parabolic motion. Used velocity 100 m/s, 30 degree when thraw the ball, the time needed to finish the path is 5.10204 s and the maximum height is 312.5 m this result is same with the calculation using equation 4 and 5. The last simulation, simulation on swing pendulum is shown by Figure 6. Swing pendulum simulation is done by inputing the degree formed while the ball swings.

Figure 4. The interface of non-uniform linear motion simulation menu

Figure 5. The interface of parabolic motion simulation menu

Figure 6. Interface of swing pendulum simulation menu

6. RESULT

The motion simulation makes an error in calculating the time, but the error is not significant, 1-2 seconds different than the ideal calculation. It is caused by the ordering of time checking in the program. Therefore, this CAL can help students to learn motion more interesting and clear. Using questionaires result, we can conclude that they learn better using our CAL than using conventional method.

7. REFERENCES

[1] Hussin, Supyan. 2007 (n,d). Pengajaran-pemelajaran bahasa. March 28, 2007. http://www.ftsm.ukm.my

[2] Koesnandar,Ade.2007 (n.d). Guru dan media pembelajaran. March 28, 2007. http://www.smuhayog.sch.id/in/module

[3] Hidayat, Dedi. 1998. Prinsip-prinsip Fisika. Jakarta: Yudhistira

[4] Kanginan, Marthen. 2004. Fisika untuk SMA kelas XI. Jakarta:Erlangga.

[5] Weisstein, Eric. 1996. Simple pendulum. March 12, 2007. http://scienceworld.wolfram.com/physics/topics/ Pendulum Physics Simulation.html

[6] Weisstein, Eric. 1996. Double pendulum. March 11, 2007. http://scienceworld.wolfram.com/physics/topics/Pendula.html

[7] Winkel,W.S. 1996. Psikologi pengajaran. Jakarta: Grasindo.

[8] Makmun, Abin Syamsuddin. 1996. Psikologi kependidikan perangkat sistem pengajaran modul. Bandung: PT Remaja Rosdakarya.

University Timetabling Problems with Customizable Constraints Using Particle Swarm Optimization Method

Paulus Mudjihartono
Informatics, Faculty of Industrial Technology
Atma Jaya Yogyakarta University, Indonesia
paul235@mail.uajy.ac.id

Wahyu Triadi Gunawan
Informatics, Faculty of Industrial Technology
Atma Jaya Yogyakarta University, Indonesia
wahyu.triadi.gunawan@gmail.com

The Jin Ai
Informatics, Faculty of Industrial Technology
Atma Jaya Yogyakarta University, Indonesia
jinai@mail.uajy.ac.id

ABSTRACT
Building a timetable or academic schedule is one of the duties a faculty officer regularly does. There are huge numbers of combination a timetable could be built, but one or a few of them is desirable. A good timetable should accommodate constraints that might be emerged due to some interests. Constraints to be considered are (1) the nature of the timetable, (2) university requirements and (3) lecturers' preferences. It is granted that there are no calculated solutions violate the nature of the timetable. It is done by forceful action in the making of solution. To extend the flexibility of the software, it is proposed to make constraints customizable. The violation of university requirements and lecturers' preferences could exist in some solutions, but of course these solutions are considered not optimized and hence not chosen. Finding the best (or at least good enough) timetable could be considered as a non-linear problem since it is solved in various ways. Particle Swarm Optimization (PSO) is then proposed to solve this problem. PSO objective function is to minimize the value of the timetable. Every violation of the constraints will make the objective function burden some penalties. To put the problem into the PSO domain, it needs problem mapping. This paper explains how this timetable problem could be mapped onto PSO, and solved satisfactorily.

Keyword
timetable, customizable constraints, problem mapping, PSO.

1. INTRODUCTION
University Timetabling Problem is one of the changeling yet interesting problem to be solved. This paper shows the works based on our university timetabling mechanism. Lecture is simply viewed as combination of lecturer, subject and class. One subject can be divided into many classes, and we will call it lecture. One class is taught by one lecturer. Session is timeslot with fixed length during which lecture is taught. There are at most five sessions in a day. Lecture could be defined as a combination of subject, class, and lecturer which occupies classroom and session. The undergraduate program accomplished within eight semesters normally. It means there are eight staging-semesters under which multiple relevant subjects to be placed. Each subject goes into one certain staging-semester. Subjects taught in 14 weeks in a semester. The proposed timetable is intended to be valid for one semester. The creation process of complete timetable is then simply viewed as inserting lectures into cells of sessions and classrooms until no more lectures left (see Table 1). Cell is viewed as a smallest unit container with fixed position and sequence. A cell is occupied by lecture that taught by one lecturer in a certain classroom at a certain session in a certain day.

Table 1. Template of complete timetable

		classroom1	classroom2	classroom3	……..
Monday	session1	lecture1			
	session2	lecture2	lecture3		
	session3				
	session4				
	session5			lecture 4	
Tuesday	session1				
	session2				lecture n
	session3				
	session4				
	session5				……..
…….					

Timetable creation process is subjected to several constraints. There are three types of constraints which is called the **nature** of the timetable, the university **requirements** and the lecturers' **preferences**. The nature of the timetable is some intrinsic characteristics any timetable would have. The university requirements are some mandatory conditions from the university regulation that any timetable would follow. The lecturers' preferences are some choices made by lecturers any timetable would provide.

There are various solutions can be explored from such a problem. Some solutions would be candidate solutions that fit the constraints. By this step, particle swarm optimization (PSO) takes place. Reference [7] explains that PSO is a population based approach using a set of candidate solutions, called particles, which move within the search space. The trajectory followed by each particle is guided by its own experience as well as by its interaction with other particles. The algorithm is explained in [2], [3] and [5]. A solution could be viewed as a particle that has an objective value. Group of solution make up swarm. Changing the position of cell the lectures are placed, or the order of how the lectures are placed is supposed to mean changing the objective value of the particle. It is wise to notice that some terms are exchangeable (only different in context) such as particle, solution and timetable.

In order to get precise model of the problem, a particular mechanism for mapping the problem into a PSO problem is then proposed. Such mechanism will be explained in chapter four.

2. RELATED WORKS

Many scientific works try to solve this problem in various methods. One approach is using graph coloring in same university as we do research [4]. The result was still unsatisfactorily because of its lack of customization of constraints. Another method being proposed was taboo search [1]. This paper assumed a uniform timeslot and fit classroom for lectures. The paper proposes approach with 'pillar' in it to generate timetable. A pillar is an aggregate sessions. This is used to construct simpler model than weekly-based lectures. Weekly-based lectures are considered complex because of some lectures have different nature of how to be taught. The result is good nevertheless some suggestions are taken into consideration later. One of the suggestions is to customize the constraints. Another scientific work [6] had just prepared the data warehouse to be explored. The data warehouse is useful when we want to dig some more information about anything not come yet. Timetable could be generated from some data already clean and consistent.

3. CONSTRAINTS TO BE CONSIDERED

There are three types of constraints: **nature** of the timetable, university **requirement**, and lecturers' **preference**. The nature of the timetable is as follows: (1) no two or more lecturers teach in the same session in a day, and (2) a lecture can only occupy one classroom at a certain session. University requirements to be satisfied are as follows: (1) lectures with one class only and come from same staging-semester could not be placed altogether in the same session in a certain day, (2) no more than three lectures that come from same staging-semester are taught in one day, (3) a lecturer could not teach more than three times a day (4) a lecturer teaches at least four days a week. Preferences to be considered are as follows: (1) preferred sessions and days the lecturer wants to teach, (2) prevented sessions and days the lecturer does not want to teach, (3) how frequent (how many sessions) the lecturer wants to teach in a day. Usually we do not change the nature of the timetable, but we do change the requirements and preferences as requested. That is why it's called customizable constraints.

4. PROBLEM MAPPING

How this timetable problem puts into PSO method is a separated action from PSO itself. A particular mapping mechanism is required to set timetable problem into PSO method.

The followings are the problem mapping:

(1) A certain **timetable** is considered as one solution, which is represented by a multi-dimensional **particle** in the PSO method. Particles are the agent for guiding the solution searching process.

(2) A **cell** of sessions-classrooms is represented by a particle **dimension**. If there are n cells of sessions-classrooms, then there would be n-dimension of searching space. Cell-1 is represented by dimension-1; cell-2 is represented by dimension-2; and so on.

(3) The **sequence** of how lectures to be inserted into cell of sessions-classrooms is represented by the position value of a particle at each dimension.

(4) The **creating** of a new timetable solution is represented by the **moving** of a particle into a new position. PSO mechanism will lead particles moving towards a better and better solution.

(5) The **objective function** is how to minimize the value of a timetable. The value of a timetable also incorporates a **penalty** for measuring how far the problem constraints are violated. It will eventually find solutions which are free from constraints violation at the end of PSO iteration process.

The mechanism of how a particle is transformed into timetable is as follows:

(1) Label each lecture with unique integer. Call the number as the ID of the lecture. Never change this ID number.

(2) Label each cell of sessions-classrooms with unique integer. Call the number as ID of the cell. Never change this ID number.

(3) Make sure the number of cells (C) is equal or greater than the number of lectures (L). In case of C is greater than L then make some dummy lectures so that the both number are equal. In case the C is less than L then the solution is impossible, unless more classrooms are added.

(4) Mark each lecture with unique real number which is corresponding with the value of each dimension of a particle position. Call the number as index of the lecture.

(5) Sort lectures according to their indices ascending then insert them into timetable starting from lecture with smallest index until all lectures completely inserted into timetable, i.e. lecture with the smallest index to be inserted at the first cell, lecture with the second smallest index to be inserted at the second cell, etc., lecture with the largest index to be inserted at the last cell. (see Figure 1)

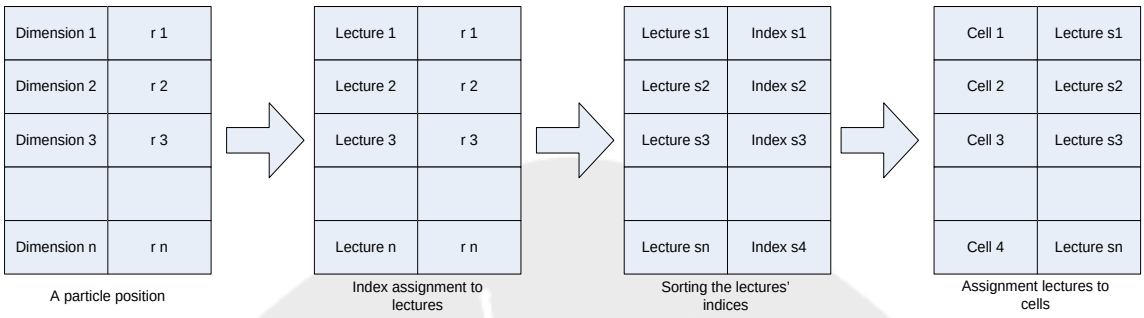


Figure 1. How a particle is transformed into timetable

The mechanism of how a particle is moving to find solution is as follows:

(1) Initialization: generate random position of *m* particles. Each particle consists of *n*-dimensional particles

(2) Transform each particle into timetable (see Figure 1). Then calculate the objective function of each timetable, assign the particle objective value equals to the corresponding timetable's objective.

(3) Use PSO mechanism to update personal best and global best position. Then, move the particles into a new position by updating first theirs velocity based on the personal and global best position.

(4) Return back to Step (2) until several iterations needed.

(5) Transform the global best position as the best timetable found by the iterations using the procedure described in Figure 1. It may not serve as the optimal timetable solution, but it is a good enough timetable solution.

Class Diagram

The implementation of idea of PSO is simply formed as a class diagram seen below. The PSO package encompasses all classes that make up the PSO method works. Some classes are omitted in order to stress the design of scheduler and its mapping into PSO. See Figure 2.

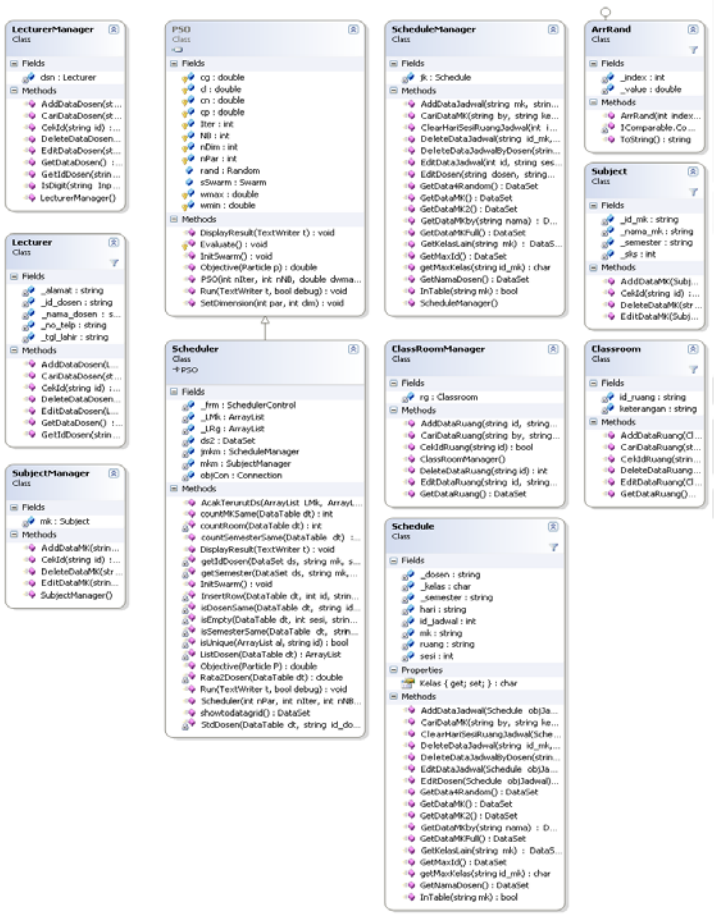


Figure 2. Class diagram PSO and its applications

5. RESULT

The result shows that there are solutions for some constraints. Some others do not have any optimized solution. More constraints to be taken care, more likely solutions not optimized. Obviously lecturers' preferences could violate the university requirements as a test entry (of course, this is not going to happen in real cases). In this case the solution will definitely not be optimized. In our test that handles constraints from 15 lecturers' preferences and all university requirements (lecturers teach 3 days a week minimum, no more that two lectures with same categorized-semester taught in same session) failed to be optimized. It is because of no more acceptable cells of sessions-classrooms available

Table 2 shows the instance of timetable with prevented session-5 in all days. It takes two rooms to hold the all lectures offered. In order to confirm some aspects, the info lied in the cell could be used. It is formatted: Subject, Class, Lecturer Code, and Staging Semester. It is confirmed that the nature of the timetable, university requirements and lecturers' preference remain fulfilled.

Table 2. Instance of timetable with last session avoidance

TIME TABLE			
Day	**Ses**	**Room 3216**	**Room 3217**
Monday	1	Mathematical Logic, A, PA, 1	Data Mining Techniques, A, ERN, 7
	2	Information System, A, AJS, 4	Web Programming, A, KA, 3
	3	Statistics, B, PA, 2	Algorithm and Programming, C, IW, 1
	4	Digital System, A, EDJ, 2	Database, A, IW, 4
Tuesday	1	Intro. Information Technology, A, KA, 1	Network Analysis and Design. A, YSP, 7
	2	Algorithm and Programming, B, PM, 1	Artificial Intelligent, A, SYT, 3
	3	Statistics, A, ERN, 2	Electronic Physics, A, EDJ, 1
	4	Calculus, A, FSP, 1	Advanced Data Structures, A, PM, 3
Wednesday	1	Corp. Mgmt Inf. Systems, A, YSP, 7	Database, C, EDU, 4
	2	Algorithm and Programming, A, EDU, 1	Informatics Economy, A, YSP, 5
	3	Calculus, C, PA, 1	Modeling and Simulation, A, SYT, 7
	4	Computer Graphics, A, SYT, 5	Advanced Computation, C, PRN, 3
Thursday	1	Advanced Computation, A, FSP, 3	Database, B, IW, 4
	2	Algorithm and Programming, D, FSR, 1	Data Structure, A, EDU, 2
	3	Intelligent Decision Systems, A, PM, 7	Computer Architecture, A, EDJ, 3
	4	Image Processing, A, BYD, 7	Intro. Mobile & Wireless Systems, A, TS, 4
Friday	1	Mathematical Logic, B, ERN, 1	Operating Systems, A, KA, 3
	2	Object Oriented Programming, A, BLS, 4	Optimization Techniques, A, FSP, 7
	3	Operation Research, A, PM, 7	Calculus, B, PA, 1
	4	Inf. System Strategic Planning, A, BLS, 7	Advanced Computation, B, AJS, 3

6. REFERENCE

[1] Adriane Mieke et al, "Tackling the University Course Timetabling Problem with an Aggregation Approach", PATAT06 Proceedings, 2006

[2] Ai The Jin & Kachitvichyanukul Voratas, "A Particle Swarm Optimization for the Vehicle Routing Problem with Simultaneous Pickup and Deleivery", Computers and Operations Research Elsevier Journal, 2009.

[3] Clerc Maurice, Particle Swarm Optimization, ISTE Ltd, 2006.

[4] Dewi Findra Kartika Sari, Schedule Generation Using Coloring Graph, Thesis, UAJY, Yogyakarta, 2006.

[5] Engelbrecht Andries P, Computational Intelligence An Introduction, Second Edition, John Wiley and Sons Ltd, 2007.

[6] Nadia, Knowledge Discovery In Academic Data With Association Rules Using C #, Thesis, UAJY, Yogyakarta, 2005.

[7] Tsou, Ching-Shih et al, "Using Croeding Distance to Improve Multi Objective PSO with Local Search", Swarm Intelligence: Focus on Ant and Particle Swarm Optimization Proceedings, 2007

A Design of Multidimensional Database for Content-based Television Video Commercial Mining

Yaya Heryadi
Binus International,
Binus Business School
yayaheryadi@binus.edu

Yudho Giri Sucahyo
Fakultas Ilmu Komputer,
Universitas Indonesia
yudho@cs.ui.ac.id

Aniati Murni Arymurthy
Fakultas Ilmu Komputer,
Universitas Indonesia
aniati@cs.ui.ac.id

ABSTRACT
The growing television and advertising industries have led to the tremendous amount of television video commercial (TVC) data stored in many data repositories of many organizations. TV broadcasting, advertising, production houses, and marketing research agencies are to name of a few. These companies constantly use information from TVC data for various purposes such as: business intelligence, advertisement tracking, copyright control, etc. The previous techniques to manage image database is not based on visual features but on textual annotation of images. Image database is indexed by textual description, i.e. keywords, captions, time of creation, etc. to facilitate queries [1,2]. However, the fast growing TVC databases have made manual annotation of images become expensive tasks.
With the advance of video warehouse technology, it is expected that a large volume of TVC database can be organized and analyzed to produce previously hidden information for decision making process. Although there are a vast number of published works in video mining, little has been said about video mining on TVC data warehouse. For that reason, this paper proposes a multidimensional database as the first step in TVC video mining. The difference of this multidimensional database design from other related works is the focus given to both key low-level visual features and metadata for TVC video mining.

Keywords
Multidimensional database, TVC data warehouse.

1. INTRODUCTION
The growing television and advertising industries have led to the tremendous amount of video data of TV program and television video commercial (TVC or TV Ad) to be stored in data repositories for further analysis of many organizations. TV broadcasting, advertising, production houses, and marketing research agencies constantly use information from TVC data for various purposes such as: business intelligence, advertisement tracking, copyright control, etc. Unfortunately, information needed for such decision making process in many organizations is only supported by textual metadata information system [3].
The current rapid increase of video data volume has made manual annotation of images for information-retrieval or information extraction become expensive tasks. With the advance of video warehouse technology, it is expected that a large volume of TVC database can be organized and analyzed efficiently. Features extracted from video content are expected to give additional value to knowledge mining to produce previously hidden information for decision making process.
Video data is a particular type of multimedia data. The later covers audio data, image data, video data, sequence data, and hypertext data which contain text, text markups, and linkages [2]. With that view, technology for video data warehouse and video mining can adapt multimedia data warehouse and multimedia mining technology.
Video data warehouse has received increasing attention following successful application of multimedia data warehouse in many areas such as: multimedia mining [4], multimedia document sharing and reused [3], spatial image mining [5, 6], medical data analysis [7], and color image retrieval [8]. Although there is a plethora of reports on video data warehouse, little attention has been given to application of this technology for multidimensional analysis of TVC data.
For that reason, the purpose of this paper is to propose a design of TVC multidimensional database as the first step in multidimensional analysis of TVC data such as video mining. The difference of the proposed design from related works is the focus given to key low-level visual features as input for many potential video mining objectives.
The remaining of this paper is structured as follows. Section 2 describes previous works. Section 3 presents the proposed multidimensional database structure. Section 4 describes data mining on TVC data cube and TVC data warehouse. Section 5 explains some TVC data mining examples followed by Section 6 as conclusion.

2. PREVIOUS WORKS
TVC is defined as "*a form of advertising in which goods, services, and ideas are promoted via the medium of television*" [9]. Following the prominent definition of data warehouse in [10], TVC data warehouse can be defined as: a subject-oriented, integrated, non-volatile, and time variant collection of TVC data in support of management's decisions process.
The key features of TVC data warehouse are:
1) Subject-oriented: The TVC data warehouse is organized around television video commercial (TVC) including: a number of visual low-level features, promoted product/service, targeted customers, clients, and airing time.
2) Integrated: The TVC data warehouse is constructed to facilitate multiple heterogeneous source and format of TVC data.

3) Nonvolatile: TVC data are permanently stored in the data warehouse which only support initial loading data and access of data functionalities.
4) Time-invariant: The TVC data are stored to provide historical information.

The advance in video data warehouse technology is highly influenced by the advance of multimedia data warehouse technologies. Many technologies applied to video data warehouse currently are adapted from previous research on multimedia data warehouse. Among those prominent works in multimedia data warehouse, Zaiane *et.al* in [4] studied a multidimensional database model to discover web access pattern from weblog in which 12 dimensions are investigated. These dimensions, among others, are: the size of image/video, width and height of frame, color, edge orientation, time of image/video creation, format of image/video data. However, the design of the proposed multimedia database is not appropriate for TVC video mining as the proposed design neither included important metadata, i.e. Brand, Target Audience Psychographic, Target Audience Social Economic Status, etc. nor main visual features, i.e. shape and texture features.

Stefanovic *et.al* in [5] investigated a spatial data cube for spatial data mining with the following dimensions: nonspatial dimension (i.e. temperature and precipitation); spatial-to-nonspatial dimension (i.e. state); and spatial-to-spatial dimension (i.e. equi-temperature region). This multidimensional model is not suitable for video mining as temporal attributes of video data are not included.

You and Liu in [8] investigated multidimensional data model for image retrieval with the following dimensions: interesting points, global color histogram, color moments, mean values of wavelet coefficients in different directions (global, vertical, horizontal, and diagonal). Despite inclusion of some low-level visual features in the model, the lack of important features such as shape and texture causes the model provide only limited support to video analysis.

Following the advance of multimedia mining technologies, video mining have been focusing on the following objectives:

1) Special pattern detection [11,12,13,14,15], which detects some predefined special patterns;
2) Association mining [14,16], which identifies association among video units;
3) Video clustering and classification [17,18,19], which clusters and classifies video units into different categories.

Such video mining applications are potentially applicable to TVC database.

In developing conceptual model of a TVC data warehouse, a multidimensional model as described in [2] is used. The model comprises of five key components as follow.

First, facts and dimensions:

1) Fact represents "*atomic information elements in a multidimensional database which consists of quantifying values stored in measures and a qualifying context determined through dimension level*" [20].
2) Dimension represents "*the perspectives or entities with respect to which an organization wants to keep records*" [2]. Dimensions together with their respective concept hierarchies allow to build a multi-dimensional data cube which is used to aggregate the values for all attributes in each dimension domain.

Candidate for data dimensions are image attributes which have been used for content-based retrieval including: color, shape, and texture [2].

Second, data cube which represents various data view in multiple dimensions.

Third, concept hierarchy to facilitate data abstraction along any given dimension such as drilling-down and rolling-up.

Fourth, measure of a data cube which represents "*a numerical function that can be evaluated at each point in the data cube space by aggregating the data corresponding to the respective dimension-value pairs defining the given point*" [2].

Finally, data schema to represent a TVC multidimensional database which are: star, snowflake, and fact constellation (galaxy) data schema [2].

Following design methodology proposed by [2], some consideration in designing TVC multidimensional database is as follow:

1) Business process to be addressed is typical TVC video mining performs by advertising and television broadcasting organizations. TVC video mining under considerations is: special pattern detection, association mining and video clustering and classification.
2) The grain of the business process is key-frame level operation in which atomic level of data will be stored in the fact table of TVC data warehouse.
3) The selected dimensions and measures are those which support TVC video mining objectives under consideration.

3. PROPOSED MULTIDIMENSIONAL DATABASE STRUCTURE

Based on the data warehouse design methodology described in [2], the proposed TVC data warehouse is designed as a single repository of TVC data.

Information stored in the TVC data warehouse represents image content of each key-frame of TVC data and global TVC information.

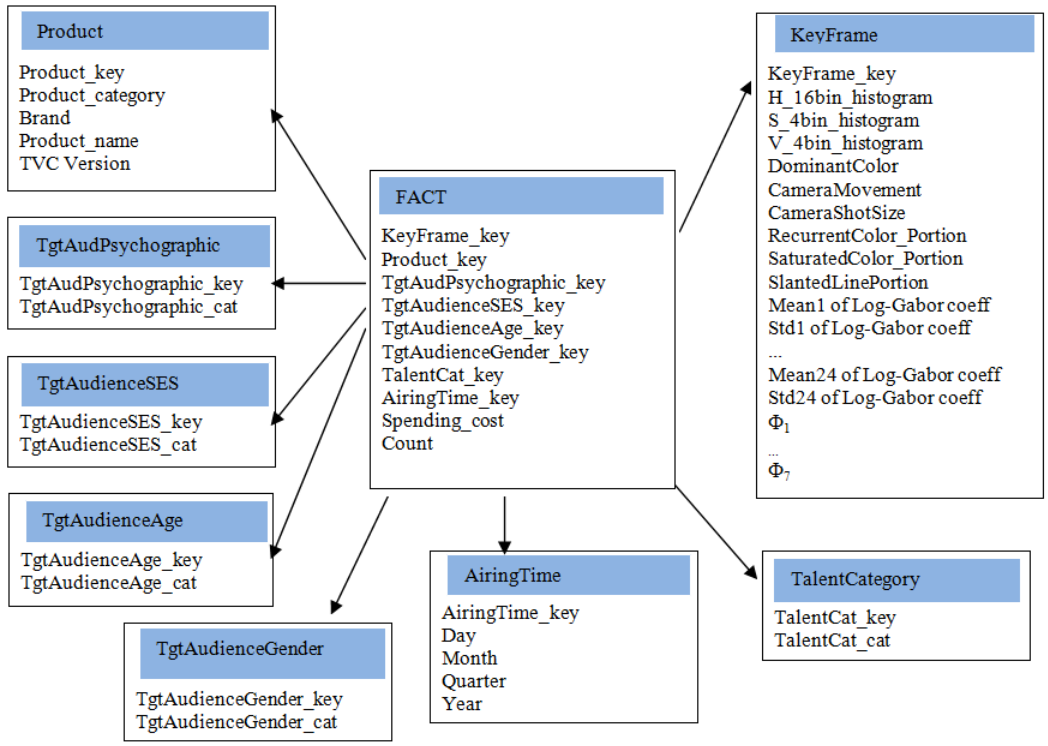


Figure 1. Star schema of the proposed TVC data warehouse

Based on the previous works, the selected low-level visual features are as follow:

1) Color is described by normalized HSV color-space histogram which is described by the following attributes: H 16 bin, S 4 bin, and V 4 bin histograms; and dominant color. The Color dimension has been explored in [14]; while, dominant color has been studied in [21, 15].
2) Camera movement whose value can be: pan left-slow, pan left-medium, pan left-fast, pan right-slow, pan right-medium, pan right-fast, zoom in, zoom out-slow, zoom out-medium, zoom out-fast, and still. The importance of this dimension has been presented in [14].
3) Semiotic features is described by attributes: portion of recurrent color, portion of saturated color, and portion of slanted lines (lines whose slope is neither horizontal nor vertical). This dimension is based on the work reported in [22].
4) Camera shot size whose value can be: loose shot, medium shot and tight shot. Selection of this dimension is based on research reported in [23].
5) Texture is represented by 24 values of mean and standard deviation of Log-Gabor transform coefficient with 4 scale levels and 6 orientations as proposed by [24].
6) Shape is described by 7 invariant moments: ϕ_1, ϕ_2, ϕ_3, ϕ_4, ϕ_5, ϕ_6 and ϕ_7 as described in [25].

In addition to those image content-based features, some additional dimensions are added for TVC mining. The main dimension which characterize TVC are: (i) Product, (ii) Airing Time, and (iii) Talent Category. Following the features proposed in [26] for segmenting advertisement audience, the proposed multidimensional database design has included the following dimensions: (i) Target Audience Psychographic, (ii) Target Audience Social Economic Status (SES), (iii) Target Audience Age, and (iv) Target Audience Gender. The fact and dimensions of multidimensional database are illustrated in Figure 1.

4. DATA MINING ON TVC DATA CUBE AND TVC DATA WAREHOUSE

OLAP and TVC data cube provide analysis environment which is very useful for extracting hidden knowledge from the TVC data warehouse. The key OLAP and data mining operations are as follow:

1) The drill-down operation navigates from general to specific concept. A drill-down along product hierarchy, for example, presenting the number of TVC grouped by version from the number of TVC grouped by Product name. The roll-up is the reverse operation of drill-down. It navigates from specific to general.
2) The slice operation performs a selection on one dimension of the given cube. For example, the selection of Month on the AiringTime dimension. The dice operation performs a selection on two or more dimensions of the given cube. For example, dicing the data cube on two dimensions: Product and AiringTime dimensions.
3) The main video mining functions as described in [2] are: TVC characterization, class comparison, association, and classification.
4) Characterizing TVC in the data warehouse is to find rules that summarize general characteristics of the TVC. For example: TVC for a particular Product Category (i.e. Cosmetics & Decorative), particular Brand (i.e. Kao, Blue

Bland) and particular Version (i.e. Fasting) can be summarized by a characteristic rule.

5) Class comparison of TVC in the data warehouse is to find discriminant rules which discriminating between different classes of TVC. For example: comparison TVC of Cosmetics & Decorative and Cigarettes product categories.
6) Association rule is to find association or correlation relationship among item data in TVC data warehouse. For example, correlation relationship between visual features for each Brand.
7) Classification is to build a model for each given classes based on the features in the TVC data warehouse and generating classification rules.
8) TVC Video Copy Detection: to identify whether a given TVC originated from another TVC by means of photometric or geometric transformations [27].

Similarity between two key frames represented by multidimensional vectors can use various metrics, i.e. Euclidian Distance as described in [25].

5. TVC MINING EXAMPLE

The proposed design of multidimensional database in this paper will be used for TVC mining. The design of TVC data warehouse is expected to support the following typical descriptive inquiries:

1) Show a diagram represents TVC spending trend (in Rp) by product category and year. As an example, the typical output is illustrated in the following figure.

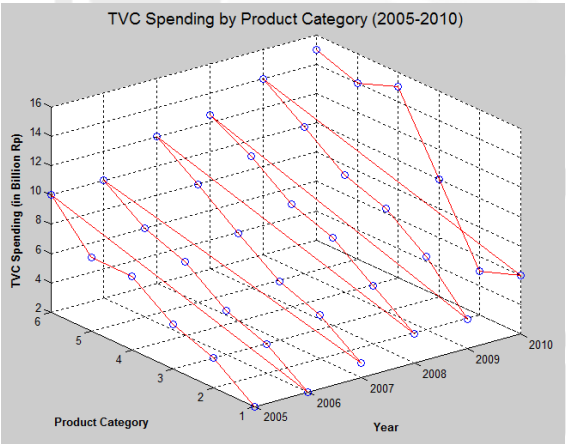


Figure 2. Diagram of TVC spending by product category

2) Show all TVCs in product category="*Toiletries*" which have visual similarity in terms of color, shape, and texture features with TVCs: Brand="*Pepsodent*", and Version="*Father&son*". Output example[1] is illustrated by *the followi*ng figure.

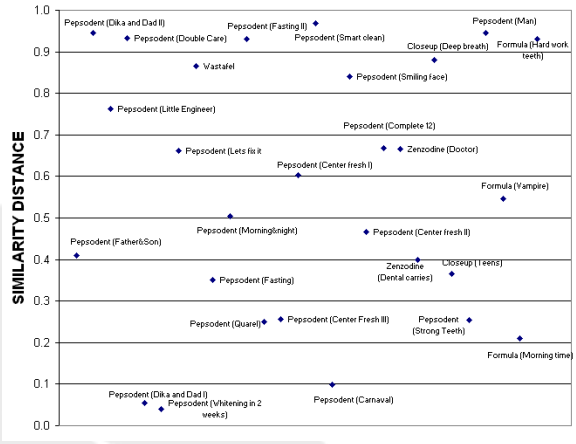


Figure 3. TVC with visual similarity with TVCs of Brand="Pepsodent", and Version= "Father&son"

3) Show all poss*ible (if any*) video copies or those video obtained from TVCs of product categ*ory="Food*", Brand="*OkeJoly*", Version="*Happy kids*".

A hypothetical example[2] is illustrated by the following figure.

Table 1. List of video copy from TVC of Product Category="Food", Brand="OkeJoly", and Version="Happy kids".

Brand	Version
JolySweet	Mom&kids
BCD Drink	Vacation

4) What kind of association that exists between TVC Brand/Version and color-based semantic as proposed by [22]. A hypothetical example[3] is illustrated by the following table.

Table 1. Association between color-based semantic and product category

Brand/Version	Color-based Semantic
Pepsodent/Father&Son	Happiness
Gudang Garam/Biker	Suspense
Molto/Baby	Relax

6. CONCLUSION

This paper has proposed a design of multidimensional database to support Data Mining from TVC data warehouse. The information stored in the TVC data warehouse combines content-based visual features and metadata for TVC video mining. The soundness of this design needs empirical data which will become the next step of this work.

7. REFERENCES

[1] Feng, D., W.C. Siu, and H.J. Zhang. "Multimedia Information Retrieval and Management: Technological Fundamentals and Applications," Springer-Verlag, New York, 2003.

[1,2,3] This figures are based on fictitious/not real data.

[2] Han, J. and M. Kamber. "Data Mining: Concepts and Techniques". Morgan Kaufmann Publishers, San Diego, CA, 2001.

[3] Ishikawa, H., K. Kubota, Y. Noguchi, K. Kato, M. Ono, N. Yoshizawa, and Y. Kanemasa. "Document Warehousing Based on Multimedia Database System," In Proceeding of the 15th International Conference on Data Engineering (ICDE'99), pp. 168-173, 1999.

[4] Zaiane, O.R., J. Han, Z.-N. Li, S.H. Chee, and J.Y. Chiang. "MultiMediaMiner: a system prototype for multimedia data mining," ACM SIGMOD Record Archive, vol. 27(2), pp. 581 – 583, 1998.

[5] Stefanovic, N., J. Han, and K. Koperski. "Object-Based Selective Materialization for Efficient Implementation of Spatial Data Cubes," IEEE Trans. on Knowledge and Data Engineering, vol. 12(6), pp. 938-958, 2000.

[6] Malinowski, E. and E. Zimanyi. "Spatial Data Warehouse: Some Solutions and Unresolved Problems," In Proceeding of the 3rd IEEE International Workshop on Database for Next-Generation Researchers, SWOD2007, pp. 1-6, 2007.

[7] Arigon, A.-M., M. Miquel and A. Tchounikine. "Multimedia data warehouses: a multiversion model and a medical application," Multimed Tools and Applications, vol. 35(1), pp. 91–108, 2007.

[8] You, J. and J. Liu. "A Data Warehousing Approach to Colour Image Retrieval," In Proceeding of the Sixth Digital Image Computing: Techniques and Applications Conference, pp. 104-109, 2002.

[9] Wang, J., L. Duan, L. Xu, H. Lu, J. S. Jin."TV Ad Video Categorization with Probabilistic Latent Concept Learning," In Proceeding of the International Workshop on Workshop on Multimedia on Multimedia Information Retrieval, pp. 217-226, 2007.

[10] Inmon, W.H. "Building the Data Warehouse 3rd ed," John Wiley & Sons Inc., New York, 2002.

[11] Sivic, J. and A. Zisserman. "Video Data Mining using Configurations of Viewpoint Invariant regions," In Proceeding of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'04), pp. 488-495, 2004.

[12] Fleischman, M., P. Decamp, and D. Roy. "Mining Temporal Patterns of Movement for Video Content Classification," In Proceeding of the 8th ACM International Workshop on Multimedia Information Retrieval, pp. 183-192, 2006.

[13] Guler, S., W. H. Liang, and I. A. Pushee. "A Video Event Detection and Mining Framework," In Proceedings of the 2003 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'03), vol. 4, pp. 42-49 2003.

[27] Shen, H.T., J. Shao, Z. Huang and X. Zhou. "Effective and Efficient Query Processing for Video Subsequence Identification," IEEE Trans. on Knowledge and Data Engineering, vol. 21(3), pp. 321-334, 2009.

[14] Zhu, X., X. Wu., A.K. Elmagarmid, Z. Feng, and L. Wu. "Video Data Mining: Semantic Indexing and Event Detection from Association Perspective," IEEE Transactions on Knowledge and Data Engineering, vol. 17(5), pp. 1-17, 2005.

[15] Lin, L. and M.-L. Shyu. "Mining High-Level Features from Video using Associations and Correlations," In Proceeding of the IEEE International Conference on Semantic Computing, pp. 137-144, 2009.

[16] Chen, M., S.-C. Chen, M.-L. Shyu. "Hierarchical Temporal Association Mining for Video Event Detection in Video Databases," In Proceeding of IEEE 23rd International Conference on Data Engineering Workshop, pp. 137-145, 2007.

[17] Oh, J. and B. Bandi. "Multimedia data mining framework for raw video sequence," In Proceeding of Multimedi Data Mining of Knowledge Discover in Database (MDM/KDD) Workshop, pp. 1-10, 2002.

[18] Pan, J.Y. and C. Faloutsos. "GeoPlot: Spatial data mining on video libraries," In Proceeding of the 11th International Conference on information and Knowledge Management, pp. 405-412, 2002.

[19] Pan, J.Y. and C. Faloutsos. "Video Cube: a novel tool for video mining and classification," In Proceeding of the 5th International Conference on Asian Digital Libraries: People, Knowledge, and Technology, pp. 194-205, 2002

[20] Husemann, B., J. Lechtenborger, and G. Vossen. "Conceptual Data Warehouse Design" In Proceeding of the International Workshop on Design and Management of Data Warehouse (DMDW'2000), pp. 3-9, 2000.

[21] Spyrou, E., G. Tolias, P. Mylonas and Y. Avrithis. "Concept detection and keyframe extraction using a visual thesaurus," Multimedia Tools Application, vol. 41, pp. 337-373, 2009.

[22] Colombo, C., A. Del Bimbo, and P. Pala. "Retrieval of Commercials by Semantic Content: The Semiotic Perspective," Multimedia Tools and Applications, vol. 13, pp. 93-118, 2001.

[23] Matsuo, Y., M. Amano, K. Uehara."Mining Video Editing Rules in Video Streams," ACM Multimedia, pp. 255-258, 2002.

[24] Field, D. J."Relations Between the Statistics of Natural Images and the Response Properties of Cortical Cells," Journal of The Optical Society of America A, vol. 4 (12), pp. 2379-2394, 1987.

[25] Gonzalez, R.C., R.E. Woods and S.L. Eddins. "Digital Image Processing using Matlab," Pearson Prentice-Hall, Upper Saddle River, NJ, 2004.

[26] Duncan, T. "Principles of Advertising & IMC," McGraw-Hill/Irwin, New York, 2005.

Applying Sound to Enhance The Comprehension of Sorting Algorithms

Lisana
Information Technology Department
The University of Surabaya
Jl. Raya Kalirungkut, Surabaya
+62 31 298 1257
lisana@ubaya.ac.id

Edwin Pramana
Information Technology Department
Sekolah Tinggi Teknik Surabaya
Jl. Ngagel Jaya Tengah 73-77, Surabaya
+62 31 502 7920
epramana@stts.edu

ABSTRACT
Computer has been successfully used as learning aid systems for the past few years. The learning systems are designed in such a way that they are easy to use and effectively convey information. Many researchers have been trying to enhance the design of the learning systems. For example, using color [8] or animation [14, 15, 16]. In this paper, we try to apply sound to enhance the learning system, more specifically to enhance the comprehension of sorting algorithms. We believe that with the combination of sound, color and animation, we can convey information more effectively in a learning system. We have conducted an experiment to test the learning system, and the study's result indicate that sound do assist students in learning.

Keywords
Algorithm Animation, Information Visualization.

1. INTRODUCTION
The audio channel is the primary medium of communication between human, however, it is rarely used between human and computer. The communication between human and computer is dominated by the visual channel. Many researches have been done in this area, but not the audio channel. The audio channel is still under-exploited. Sound is mostly used only as a flat beep to alert user of a particular event. This is an area in transition, and audio is a useful option in many design situations.

With the advances of technology, most of computer systems now have their own sound system. There are no other excuses not to explore the audio channel because of the hardware limitation. Computer systems are now capable to produce digitized sound, thus enable them to produce almost any sound needed [1].

Sounds can be interpreted at several levels [2]. For the specific function, many researchers are also interested in using sound for learning, such as teaching English by associating a particular audio signal with all verbs [3]. Moreover, some techniques that focus on sound in workstation-based interactive algorithm-animation systems had also been proposed by Brown [8].

Algorithms animation is a form of a *program visualisation*, "the use of the technology of interactive graphics and the crafts of graphic design, typography, animation, and cinematography to enhance the presentation and understanding of computer programs. Program visualisation is related to but distinct from the discipline of *visual programming* which is the use of various two-dimensional or diagrammatic in the programming process" [9]. It is concerned with illustrating the behaviour of a program by visualizing the fundamental operations of the program as it runs. Developers of algorithm animations believe that algorithm animations are useful to be used in learning how the algorithms work [12]. Some learning systems have been developed using graphics and animation in order to help explain how the algorithms work. Najork [15] proposed the use of full-fledged interactive algorithm animation that includes a rich set of libraries for creating 2D and 3D animations. In contrast, Stasko, et al [16] has proven that algorithm animations were not as helpful as was hoped. Brown stated that algorithm animation can be enhanced by using color and sound [8]. The combination of animation, color and sound can be very powerful to enhance the user interface. Sounds can effectively convey information. In this paper, we explore ways to use sound in the interface to enhance the comprehension of sorting algorithms.

2. OBJECTIVES
The main aim of this research is to apply the particular tones to some sorting learning systems, including Bubble, Selection, Insertion, Shell-Metzner, and Quick Sort, in order to enhance their user interface comprehensibility. More specifically, some sorting techniques are to be understood and studied first, followed by their implementation using Delphi. In order to apply sound to those processes, it is needed to understand about how to use the sound through the user interface.

3. USING SOUND IN USER INTERFACES
Different with visual messages where we have to see the message to understand it, audio messages are received regardless of where one is looking. This is very important when visual channel is focused elsewhere, or when the task does not require constant visual monitoring. When the amount of information to be conveyed is high, pushing the visual channel to the limits, the audio channel can also be used to carry some of the information, thereby reducing overall load.

3.1 Structuring Sound
Blattner, et al, mentioned in his article [6] that different approaches have been employed when using sound for auditory data display, on the one hand, and for messages or audio cues, on the other. This is because auditory data display has been concerned with the mapping of data points (in an *n*-dimensional space) to audio output., while the use of sound for messages or cues has been more concerned with the syntax and semantics of the audio output.

Auditory display techniques are used, for the most part, to enable the listeners to picture, in their minds, real-world objects or data. There is a lot of leeway for the interpretation of how this may be successfully achieved; human factors studies are required to understand what may actually be heard.

Auditory messages or signals were used by people before the discovery of electricity. Bells, bugles, trumpets, and drums sent information to the countryside or announced the arrival of an important person or messenger [5]. Auditory cues used to reduce visual workload were studied by Brown, Newsome and Glinert [10]. The purpose of these cues was to identify the location of groupings of letters on a screen; hence, these cues may be considered "auditory pointers" to items on the screen.

Audio messages have also been studied by Gaver [11], Blattner [4], and by Blattner, Greenberg, and Kamegai [5]. Gaver used "real-world" sounds, called *auditory icons*, to convey messages. His examples include objects colliding, breaking, and so on. The fact that Gaver used "real-world" sounds does not imply anything about the meanings of these sounds in his system, which could be an abstraction (eg. "your computer is going down").

Blattner, et al, used *earcons*, which are tones or sequences of tones, as a basis for building messages [5]. Earcons and auditory icons are synonymous terms in the sense that they both denote the auditory form of an icon or visual symbol. The distinction lies in the approach to their construction. Whereas Gaver uses the word "icon" in its original meaning (that is, highly representational images), Blattner, et al, based earcons on similarities between their auditory messages and abstract visual symbols.

The difference between these approaches is the difference between lexical, syntactic, and semantic levels of a language. Syntax is the formal composition of elements in a language. Semantic is the meaning given to the structures in the language. The lexical level refers to the properties of the base elements. The lexical level of auditory messages is concerned with the attributes of sounds: frequency, pitch, duration, loudness, etc.

Most of work done with the auditory display of data is concerned with the attributes or parameters of sound. This effort has been concentrated on the lexical level. Because earcons are not a direct translation of data into audio, they are able to use a formal syntax in the composition of messages-messages in regard to real-world sounds. Auditory icons are studied primarily on the semantic level. The syntax of auditory icons does not ordinarily arise, because such icons are sampled sounds composed of very complex waveforms and not easily parameterized. Hence, the important contribution of the work on auditory icons to the treatment of messages as a language is our deeper knowledge of the "meaning" we give to sounds.

3.2 Using Earcons and Animation

In the musical world, a short sequence of tones is called a *motive*. In the construction of earcons, a motive is used as a building block for larger groupings. The advantage of these constructions is that the musical parameters of rhythm, pitch, timbre, dynamics (loudness), and register can be easily manipulated. The motives can be combined, transformed, or inherited to form more complex structures. The motives and their compounded forms are called *earcons*. However, earcons can be any auditory message, such as real-world sound. In the implementation used in this article, the messages are information about data itself and events of the data rather than translation data into sound. One advantage of structuring audio messages in this way is that they can be used with any basic sonic unit, such as tones or sampled sounds.

Data translated directly into sound require less explanation or motivation than abstractions such as earcons. Although auditory icons that make real-world sounds usually can be recognized quickly, several experiments have shown that earcons are preferred over many other types of sonification. Jones, et al, compared earcons, auditory icons, and synthesized speech; their result showed that subjects preferred the sounds of earcons but were better able to associate auditory icons to commands [13]. Earcons was found to be an effective form of auditory communication [7].

Blattner gave a hint in applying earcons that earcons are necessarily short, because they must be learned and understood quickly [6]. Earcons were designed to take advantage of chunking mechanisms and hierarchical structures that favor retention in human memory. The tests run by Brewster, et al had no training session associated with them (the earcons were heard only once before conducting the test); in spite of this, the subjects could use them effectively.

We are experimenting with introducing earcons, tones and sampled sounds, to the users as an animation with sound. Sound is the result of movement, the interactions and mapping of data. The animation describes the data or event that the motive supports. For example, suppose our animation was swapping, the motive that forms the basis of the earcon is the sound of the swapping process.

The reason why we prefer to used tones rather than auditory icons, because in the realism of the sounds used in the auditory icons led to confusion on the part of the users. For example, a hammer hitting a table may have the same sound as a car door slamming. Moreover, the most real-world sound, auditory icons, are complex and difficult to analyse. Tones would never be confused with the real-world sounds. An added bonus is that our sounds are easier to manipulate, for example by using wave-editor. The point here is that the sounds do not have to be realistic if the user is presented with a method of determining what the sounds represent.

4. THE DESIGN PROCESS

In order to produce a user-friendly learning system design, firstly, we conducted task analysis in order to achieve an information source from which design decisions could be made, and a basis for evaluating designed systems. After conducting task analysis, we analized the result and decided the final design, shown in section 4.2.

4.1 Knowledge Analysis of Tasks (KAT)

KAT methodology can produce a complete and explicit model of tasks in the domain, and of how people carry out those tasks. It focuses design on user's tasks and goals, and the methods for achieving those goals, resulting in improved, more usable system designs.

4.1.1 Setting Up

The first step from KAT methodology is the setting up which shows the objective. The purpose of this analysis is to gain the information what the users expect when they want to learn the sorting algorithms. This must take place in an environment where there are many potential users, with typical expectations.

4.1.2 Data Collection

Data was collected from the various users, according to their education background, through interview. The reason why we chose interview technique is because this sorting learning system needs extracting rules and the interview technique takes less time. As the result, we got an initial view of the set of tasks that should be put in our learning system.

For the users whose education background are not computer literate, it's quite difficult for them to specify what they should expect when they want to learn some sorting algorithms. In contrast, for good computer knowledge users, they know well what should be in a sorting learning system. That's why, we decided that this sorting learning system is purposed specially for computer science graduate students who had taken or were taking advanced computer algorithms courses.

4.1.3 Identifying Knowledge Components

KAT is concerned with identifying a person's task knowledge in terms of actions and objects, and the structure of those objects, procedures, the task plan, task goals and subgoals.

Table 1. List of objects and actions

Objects	*Actions*
Memo1	Explain the current process which follows the current animation.
Memo2	List all occured comparisons and swappings from a particular sorting algortihm.
Sound Panel	1. Change to animation only option 2. Change to sound-animation option
Sorting algorithms Panel	Select the desired sorting algorithm.
Speed Panel	Change the animation speed.
Pause button	1. Stop the current process temporary. 2. Continue the paused process.
Stop button	Terminate the current sorting algorithm process.
Start button	Start new sorting algorithm process.

Objects and their associated actions used in carrying out the sorting tasks were identified by analyst ourselves carrying out the task. We applied this technique because we have sufficient computer knowledge, so we analysed the data collected from section 4.1.2 and we decided the objects and the actions. As the result, in the figure 1, we listed the objects and their associated actions to carry out the sorting tasks.

Technique was used in this learning system in order to identify a person's knowledge of the task plan, the sequence of carrying out routine procedures, and strategies used in the task, was *asking specific questions in the structured interview* technique.

4.1.4 Data Analysis

After data is analysed, we got a set of tasks that users expect to do whenever they want to learn the sorting algorithms, listed below:

- restart learning a particular sorting algorithm
- pause the process
- change the process speed
- stop the process
- change the sorting algorithms

4.2 Design Layout

Finally, we have developed the final sorting algorithms learning system design, shown in Figure 1.

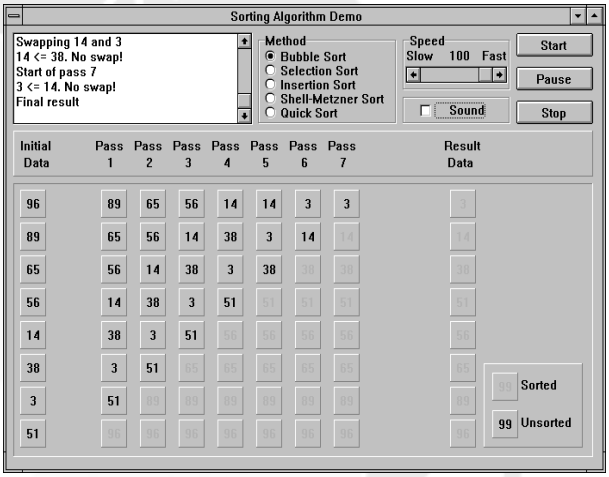


Figure 1. Screen design of a bubble sort demo

5. IMPLEMENTATION

We implemented this sorting algorithms learning system by using Delphi as the development tool and Windows as the platform of the system regarding to the final design that we had developed. The learning system is having sound capabilities and color animation.

This sorting learning system implements five sorting algorithms: Selection Sort, Bubble Sort, Insertion Sort, Shell-Metzner Sort and Quick Sort algorithms. Mouse was used as an interaction between users and the system.

For creating the sound, both sample sounds and tones were created on the computer through keyboard and the sound files were in .WAV extension. The average size of each sound files is about 10 Kb. In order to explain how the sorting algorithms work, for animation, we used eight numbers as the data filled randomly, and each will represent a particular tone in an octave according to the rank.

6. SYSTEM EVALUATION

This evaluation applied the same approach presented by Stasko et al. [16] who examined the viability of algorithm animation as a tool for learning a particular data structure and algorithm - the pairing heap data structure. They used 20 subjects, all volunteers, were computer science graduate students who had taken or were taking advanced computer algorithms courses. None of the subjects had

ever studied pairing heaps. Each subject was randomly assigned to be in one of two groups, with ten persons total per group. The first group was given textual descriptions of the pairing heap algorithm, and the second group was given the same textual descriptions supplemented by the opportunity to interact with the pairing heap animation.

The aim of the evaluation purposed in this article is to determine whether sound can be used to enhance the comprehension of sorting algorithms or not.

6.1 The Experiment

In this experiment, we used Sorting Demo system that had been implemented, specified in section 5. Before we used this system in this experiment, software testing had been conducted in order to make sure that the system has been free from any bugs. We referred this testing to as verification and validation. After everything had been fixed the experiment was started.

The computer that we used in conducting this experiment is using CPU Pentium IV, 3.0 GhZ with 512 MB RAM, with Microsoft Windows XP SP2.

Each algorithm supplies the capability for the users to change the speed of the animation, pause and stop the current sorting process, restart the current sorting algorithm, and start the new sorting algorithm. The same sequence of numbers is used for all sorting algorithms and all subjects in order to improve the validity of this experiment. We use a limited sequence of numbers so that the subjects can understand the sorting algorithms easier.

Fifty subjects, all volunteers, were participated in this experiment; they were Information Technology undergraduate students either had taken or were taking computer algorithms courses at the University of Surabaya to simulate the conditions that the system would be used. We had two groups, twenty-five subjects per group assigned randomly. The first group received textual descriptions of the five sorting algorithms with opportunity to interact with the Sort Demo system without sound, in other word with animation only, and the other group received the same textual descriptions with opportunity to interact with the complete Sort Demo system, including sound and animation. This textual description gave the information about data structure of each sorting algorithms and how each of those algorithms are implemented.

Each subject was given 45 minutes to complete the study, an initial of 20 minutes to read and understand five sorting algorithms from the given descriptions and the remaining time (to a total of 45 minutes) to interact with the system; the animation system for group one and the sound-animation system for group two.

Special for group two, firstly the subjects in this group were given 5 minutes training about all earcons which they would hear in the system, including their association (the earcons were heard only once before they interacted with the system). All the subjects in both group one and group two received some explanations about how to operate either the animation system or sound-animation system according to which group they were in. They were allowed to interact with the system in any manner they desired, e.g. pause and stop the process, restart learning a particular sort algorithm.

When the subjects had completed the 45 minutes learning session, we gave them a set of questions to test their understanding of the sorting algorithms. Neither group was allowed to use the textual description nor the system while the examination was in progress. The subjects were given a maximum of 15 minutes to work on the exam. The questions were designed as the essay-style questions, one question per each sorting algorithm so there were five questions per subject.

The sample question is organized to a major section applied to each sorting algorithms, stated below:

> Given the following numbers below, show the order of these numbers of pass one and pass two of the Bubble Sort algorithm in order to make these numbers sorted.
>
> 3 1 8 5 2

6.2 Results and Discussion

Recall that we had fifty subjects, twenty-five subjects were in each of the animation and sound-animation groups. Table 2 shows the result of the experiment. This table lists the number of correct replies for each group. The correct replies are listed in the same order as the description of the question categories described earlier in the article.

Table 2. Results of the experiment.

Sorting Algorithm	*Animation Only*	*Sound-Animation*
Bubble Sort	18	23
Selection Sort	20	21
Insertion Sort	18	17
Shell-Metzner Sort	12	16
Quick Sort	5	4

As we can see from the table that the sound-animation group performed well or better than the animation group in three kinds of questions, not in Insertion Sort and Quick Sort questions. We note here that additional of sound in this learning system can enhance the comprehension of sorting algorithms. From the result above, it seems that the sound's benefit was not too strong, but for general we can see that the subjects from sound-animation group felt confident in answering the given questions.

Another important input for our analysis was from debriefing section with the subjects in the sound-animation group. It was held after the experiment was finished. We tried to get information: whether the subject felt that the sound aided understanding the sorting algorithms. All twenty-five sound-animation subjects stated that they felt the sound assisted them in understanding the sorting algorithms, eventhough at the first time they confused in matching between watching the animation and hearing sound. The animation was often grabbed their attention. After a few minutes, they could interact with the system well and found that sound could help them in learning.

Some subjects said that after they were familiar with the sound, they could learn the sorting algorithms relaxed without concentrating too much watching on the screen. They could enjoy watching the animation while hearing the sound. One said that sound could make the sorting algorithm clear what was happening instead of watching the animation.

To get more detail information, we let the subjects who were in sound-animation group to interact again with the system without sound. Just to get information whether they felt that sound-animation system was better than animation system. All those subjects gave the same comments that sound-animation was much better than animation only.

They cited negatively that Quick Sort animation was the most terrible one. Most of the subjects did not properly understand how the sorting algorithm works. Even they could recognize the kind of sounds when Quick Sort was in its action, but they confused with the animation. The Quick Sort result was not too surprising because we know that this sorting algorithm is the most difficult one to be understood and also our Quick Sort animation was not too good. Blattner said in his article that designing an enlightening animation is a tricky psychological and perceptual challenge. At present, creating effective dynamic visualizations of computer programs is an art, not a science [5].

7. CONCLUSION

7.1 Concluding Remarks

Finally, this research have proved that applying sound do enhance the comprehension of sorting algorithms. Combining sounds, color and animation can assist students in learning and understanding sorting algorithms. We use color for encoding the state of data structure, highlighting activity, trying multiple views together, emphasizing patterns, and making an algorithm's history visible in a single static image. We use sound for reinforcing visuals, conveying patterns, replacing visuals, and signalling exceptional conditions.

Moreover, this research's result is also supported by Brown who stated that color and sounds do not merely enhance the beauty of a presentation, but also present fundamental information [8].

7.2 Further Study

We believe that sound has much more capabilities in conveying information. Maybe the sounds used in this experiment are not the best ones. We believe that sound will be more helpful when we use the right sound for a particular purpose.

8. REFERENCES

[1] Alty, J.L., Rigas, D. & Vickers, P. 1997. Using Music as a Communication Medium. CHI 1997 extended abstracts on Human Factors in Computing Systems: lloking to the future, pp. 30-31.

[2] Arons, B. & Mynatt, E. 1994. The Future of Speech and Audio in the Interface. SIGCHI, ACM, 26(4), pp. 44-48.

[3] Blattner, M.M. and Greenberg, R.M. 1992. Communicating and Learning Through Non-speech Audio. NATO ASI Series, Vol. F 76: Multimedia Interface Design in Education, pp. 133-143.

[4] Blattner, M.M., Sumikawa, D.A. & Greenberg, R.M. 1989. Earcons and Icons: Their Structure and Common Design Principles, Human Computer Interaction. 4(1), pp. 11-44.

[5] Blattner, M.M., Greenberg, R.M. & Kamegai, M. 1992. Listening To Turbulence: An Example of Scientific Audiolization. Blattner, M.M. and Dannenberg, R.B. (Ed.), Multimedia Interface Design, Addison-Wesley, Reading, Massachusetts, pp. 87-102.

[6] Blattner, M.M., Papp, A.L. & Glinert, E.P. 1994. Sonic Enhancement of Two-Dimensional Graphic Displays. Kramer, G. (Ed.), Auditory Display: Sonification, Audification, and Auditory Interfaces, Proceedings vol. XVIII, Santa Fe Institute, Addison-Wesley, Reading-MA, pp. 447-470.

[7] Brewster, S.A., Wright, P.C. & Edwards, A.D.N. 1994. A detailed investigation into the effectiveness of earcons. Kramer, G. (Ed.), Auditory Display: Sonification, Audification, and Auditory Interfaces, Proceedings vol. XVIII, Santa Fe Institute, Addison-Wesley, Reading-MA, pp. 471-498.

[8] Brown, M.H. & Hershberger, J. 1992. Color and Sound in Algorithm Animation. IEEE Computer 25(12), pp. 52-63.

[9] Brown, M.H. 1988. Perspective on Algorithm Animation. Proceedings of ACM CHI '88, pp. 33-38.

[10] Brown, M.L., Newsome, S.L. & Glinert, E.P. 1989. An Experiment into the Use of Auditory Cues to Reduce Visual Workload. Proceedings of ACM CHI '89, pp. 339-346.

[11] Gaver, W.W. 1986. Auditory Icons: Using Sound in Computer Interfaces. Human Computer Interaction, 2(2), pp. 167-177.

[12] Hundhausen, C.D. 2002. Integrating Algorithm Vizualisation Technology into an Undergraduate Algorithms Course: Ethnographic Studies of a Social Constructivist Approach. Computers & Education, v.30 n.3, pp. 237-260.

[13] Jones, S.D. & Furner, S.M. 1989. The construction of audio icons and information cues for human-computer dialogues. Contemporary Ergonomics, Proceedings of the Ergonomics Society's 1989 Annual Conference, T. Megaw (Ed.), pp. 436-441.

[14] Karavirta, V. 2009. Seamless Merging of Hypertext and Algoritm Animation. ACM Transactions on Computing Education (TOCE) v.9 issue 2, article 10.

[15] Najork, M. 2001. Web-based Algorithm Animation. Proceedings of the 38th Annual Design Automation Conference, pp. 506-511.

[16] Stasko, J., Badre, A. & Lewis, C. 1993. Do Algorithm Animations Assist Learning? An Empirical Study and Analysis. ACM INTERCHI'93, pp. 61-66.

Data Mining to Build A Pattern of Knowledge From Psychological Consultation

Sri Mulyana
Computer Science Study Program
Department of Computer Sciences and Electronics
Gadjah Mada University
Bulaksumur Yogyakarta
smulyana@ugm.ac.id

Sri Hartati
Computer Science Study Program
Department of Computer Sciences and Electronics
Gadjah Mada University
Bulaksumur Yogyakarta
shartati@ugm.ac.id

Retantyo Wardoyo
Computer Science Study Program
Department of Computer Sciences and Electronics
Gadjah Mada University
Bulaksumur Yogyakarta
rw@ugm.ac.id

Edi Winarko
Computer Science Study Program
Department of Computer Sciences and Electronics
Gadjah Mada University
Bulaksumur Yogyakarta
ewinarko@ugm.ac.id

ABSTRACT

Psychological problems require a special handling that is usually carried out by a psychologist through a consultancy. Reasoning Systems or Expert Systems for psychological counseling, it is possible to developt. This paper will describe the process of data mining to build a pattern of knowledge from psychological consultation notes that will be used to develop an expert system in the next step. The process uses Rapidminer 4.0 software with decision tree and rule learner methods. The results of the process of data mining by decision tree shows that the data converge on the type of action ass (assistance). The results of the data mining method by the rule learner managed to find a pattern of knowledge of rules 15 rules.

Keywords

Data Mining, Knowledge Discovery, Case Based Reasoning, decision tree, rule learner.

1. INTRODUCTION

One of the most current approaches to artificial intelligence involves constructing programs that function as narrowly focused experts called expert systems. Expert system is a new innovation in the capture and integrates knowledge, has the ability to duplicate the expertise of an expert in a particular field.

Psychological problems require special handling that is usually done by a psychologist through a consultancy. The more complex problems of life will create a psychological problem and needed more the role of psychological experts to solve. Expert Systems for psychological counseling, it is possible to help resolve the issue.

The development of Expert Systems requires a large data management process of psychological consultation notes. The overall knowledge of the search process is called Knowledge Discovery in Databases (KDD). Data mining is one step in the process of KDD using of a specific algorithm in the search pattern in database. This paper will explain the process of data mining to construct a pattern of knowledge from the psychological consultation notes, which will be used to develop the expert system in the next step using Case Based Reasoning techniques.

In this research, data mining process has been carried out to build a pattern based on knowledge of psychology counseling. The process uses Rapidminer 4.0 software with decision tree and rule learner methods.

2. CASE BASED REASONING

Case-Based Reasoning (CBR) has become a successful technique for knowledge-based systems in many domains. A short definition of case-based reasoning is that it is a methodology for solving problems by utilizing previous experiences (Kolodner, 1993). In case-based reasoning, a reasoner solves a new problem by noticing its similarity to one or several previously solved problems and by adapting their known solutions instead of working out a solution from scratch.

The problem-solving life cycle in a CBR system consists essentially of the following four parts (Aamodt & Plaza, 1994):

1. ***Retrieving*** similar previously experienced cases (e.g., problem–solution–outcome triples) whose problem is judged to be similar
2. ***Reusing*** the cases by copying or integrating the solutions from the cases retrieved
3. ***Revising*** or adapting the solution(s) retrieved in an attempt to solve the new problem
4. ***Retaining*** the new solution once it has been confirmed or validated

The relationship between these steps can be presented below:

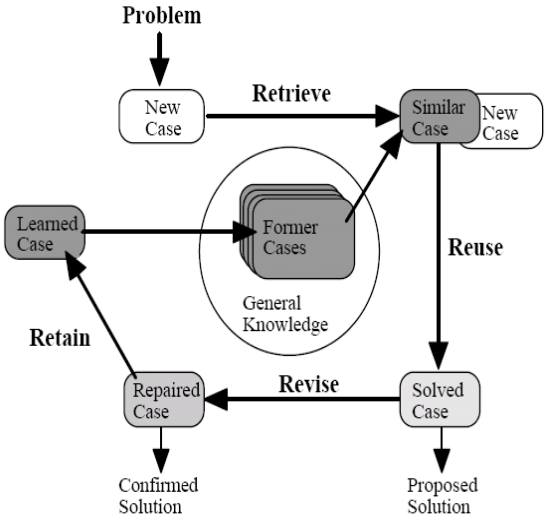


Figure 1. CBR cycle (Aamodt & Plaza, 1994)

A new problem, described as a case, is compared to the existing cases in the case base and the most similar case or case are retrieved. These cases are combined and reused to suggest a solution for the new problem. The solution proposed may be needed to be revised somewhat if it is not a valid solution. Then, the verified solution is retained by adding it as a new case to the case base or as amandements to existing cases in the case base for use in the future problem solving.

3. DATA MINING

'Data Mining' may be defined as the process of searching and analizing data in order to find implicit but potentially useful information [Frawley at all, 1991]. There are two types of data mining in its aplication (Kantardzic, 2003).

a. Predictive data mining
b. Descriptive data mining

The most popular data mining techniques and frequently used in solving the problem are (Pramudiono, 2003):

a. *Classification*
Classification is the processes to find a model or function that distinguish data classes, in order to estimate the class of an object with unknown label**.** The model can be "IF-THEN" rules, *decision tree*, mathematical formula or *neural network*.

b. *Association Rule*
An association rule is a data mining technique to find the combination of associative rule items. The importance of association rules can be identified based on two parameters y *support* and *confidence*. Support is the percentage of item combinations in the database, and confidence is the level of relations between items in association rule. The Algorithm commonly used in association rule is apriori algorithm.

c. Clustering
Clustering is a process of classifying data that are not based on a specific data class. In fact, clustering can be used to provide the labels of unknown class data. Clustering is often classified as an *unsupervised learning* method.

4. METHODOLOGY

This research is part of the whole research process in the context of preparing a dissertation, entitled "Computer Reasoning System for Expert based on case" with a case study on the psychological consultation. The purpose of this research is to build a knowledge base about the psychological consultation problem. The research methods are designed as follows:

a. Data Collection
This stage is for collecting data that contain records of clinical psychological consultation from the psychological consultant. This is a very important and strategic stage, because the computer reasoning system is developed based on the knowledge base obtained from the data.

b. Data Normalization
Data normalization covering several processes as follows :
- *Data Cleaning:* a process for removing noise and inconsistent data.
- *Data integration*: the process of collecting data from various sources into a data base called data warehouse
- *Data selection and transformation*: a process for transformting data into a form suitable for data mining and then making the selection of data.

The purpose of this normalization is to change and choose the data that can be implemented on data mining algorithms.

c. Data Mining Process
The purpose of this process is to build knowledge related to psychological consultation. This process will utilize Rapidminer 4.0 software.

d. Knowledge Verification
Consultation with the psychological consultant about the knowledge that are generated from data mining process.

5. RESULTS AND DISCUSSION

The data records that have been normalized was used for data mining process. Based on the data mining process, the knowledge generated in the form of production rules (IF .. THEN) contains attribute of age, sex, class, medical diagnosis, and psychological diagnosis for antecedent, and attribute treatment for concequent. The process uses Rapidminer 4.0 software with decision tree and rule learner methods. The process for detail is as follows :

a. Data input for Rapidminer
The form of file *.aml was used for Data input. The file was saved as <label></label> for the consequent and <attribute></attribute> for the antecedent. The file will call other file which form *.data acording to the sequence.
Figure-2 is an example for the *.aml file.

```
<attributeset default_source="penelitian.data">
 <attribute
   name        = "umur"
   sourcecol   = "1"
   valuetype   = "integer"
 />
 <attribute
   name        = "jenis_kelamin"
   sourcecol   = "2"
   valuetype   = "nominal">
```

```
        <value>L</value>
        <value>p</value>
  </attribute>
  <attribute
    name        = "kelas"
    sourcecol   = "3"
    valuetype   = "nominal">
        <value>III</value>
        <value>II</value>
        <value>I</value>
    <value>VIP</value>
  </attribute>
  <attribute
    name        = "dx_medik"
    sourcecol   = "4"
    valuetype   = "nominal"
  />
  <attribute
    name        = "dx_psy"
    sourcecol   = "5"
    valuetype   = "nominal"
  />
  <label
     name     = "jenis_tindakan"
     sourcecol = "6"
     valuetype = "nominal">
  </label>
</attributeset>
```

Figure 2. Sourcecode Penelitian.aml

Figure-3 is an example for the *.data file.

```
30,L,III,conipUTHXIIFr.a,stress,ass-kons
13,L,III,Scoliosis,stress,ass
38,L,III,paraparesis,stress,ass-kons
50,L,III,RowstFractULIFr.B,Denial,ass-kons
27,L,III,SponditotitosisULI,cemas,ass
43,p,II,LBHSusp.Hnp,stress,ass
36,L,III,paraplegia,stress,ass
31,L,III,paraplegia,adjustime,kons
6,p,III,Scoliosis,normal,kons
...
```

Figure 3. Sourcecode Penelitian.data

b. Data Mining Process

1) *Decision Tree Method*

The output of data mining process by decision tree method could be shown as follow :

```
ass {ass-kons=61, ass=196, kons=16,
psiter-psitest=1, ass-psiter=19, ass-
kons-psiter=2, kons-psiter=1, psiter=2,
ass-psitest=2, Ass-Psiter=37, Psiter=40,
Ass-Kons=49, Ass-Kons-Psiter=3, Ass=49,
PsiDasar-Konseling=5, PsiDasar-
Psikoterapi=5, PsiDasar=1,
Psikoterapi=4, PlayTerapi-Konseling=1,
PsiDasar-Konseling-Psikoterapi=1, Ass-
PsiTest=2,
Kons=3}
```

The results of the process of data mining by decision tree shows that the data converge on the type of action ass (assistance), because most of the records (196 records) have the kind of ass (assistance) action although antecedents diverse.

2) *Rule Learner Method*

This method will extract the rule from penelitian.aml data input. Numerical data was divided in the range form using *FrequencyDiscretization*. The feature of age was divided into 3 ranges. Then *number of bins=3* and sample ratio = 1.0 was chosed. It means that all records were included in the determination of the rule.

This process generates the rules shown in Figure-4.

Figure-4 show that :

- There are 15 rules.
- The first rule : if dx_psy = normal then ass (5/51/1/0/1/0/0/0/0/0/0/0/0/0/0) mean that if psycological diagnosis = normal then the treatment = ass. The sequence numbers show the number of records that support and unsupport the rule. There are 51 supported data and 7 unsupported data for the rule.
- "*correct*: 277 *out of* 500 *training examples"* mean that there are 277 data from 500 data supporting all of the rules. So the significance of this rule was 55 % (277/500 * 100 %).

```
1. if dx_psy = normal then ass
(5/51/1/0/1/0/0/0/0/0/0/0/0/0/0)
2. if dx_psy = cemas then ass
(66/149/16/0/23/2/0/2/1/0/2/0/0/0/0/1)
3. if dx_psy = MoodDisorder then ass
(18/23/0/0/1/1/0/0/0/0/0/0/0/0/0)
4. if kelas = III then psiter
(14/15/1/0/21/2/1/31/1/0/0/0/2/0/0/0)
5. if jenis_kelamin = L then ass-psiter
(3/2/1/1/6/0/0/3/0/0/1/1/1/0/0/1)
6. if kelas = II then psiter  (0/2/0/0/2/0/0/5/0/0/0/0/0/0/0/0)
7. if dx_psy=anxietas then psiDasar-konseling
(2/2/0/0/1/0/0/0/0/4/2/0/0/1 /0/0)
8. if dx_medik = Fr.InterochanterFemur then ass-kons
(1/0/0/0/0/0/0/0/0/0/ 0/0/0/0/0/0/0)
9. if dx_medik = MultipleFraktur then psiter
(0/0/0/0/0/0/0/1/0/0/0/0/0/0/ 0/0)
10. if umur = range1 then ass-psiter
(0/0/0/0/1/0/0/0/0/0/0/0/0/0/0/0)
11. if dx_medik = Fr.compresiVertebrae then psikoterapi
(0/0/0/0/0/0/0/0/0/0/0 /0/1/0/0/0)
12. if dx_medik = Multifractur then psiDasar-konseling-
psikoterapi  (0/0/0 /0/0/ 0/0/0/0/0/0/0/0/0/1/0)
13. if dx_medik = Fr.collFemur then psiDasar-konseling
(0/0/0/0/0/0/0/0/0 /1/0/ 0/0/0/0/0)
14. if umur = range2 then ass
(0/1/0/0/0/0/0/0/0/0/0/0/0/0/0/0)
15. else ass-kons  (1/0/0/0/0/0/0/0/0/0/0/0/0/0/0)
correct: 277 out of 500 training examples.
```

Figure 4. Generated rules

6. CONCLUSIONS

Data mining process by decision tree method can't generates pattern of knowledge. On the other hand, data mining process by the rule learner method generated 15 rules and the degree of significant was 55 % .

7. REFFERENCES

[1] Aamodt, A., 1994, *Case-based reasoning: Foundation issues*. *AICOM* **7** 39-59

[2] Frawley W.J, Shapiro G.P., Matheus C.J., 1991, *'Knowledge Discovery in Database : An Overviw'*, in Knowledge Discovery in Database, AAAI Press, Menlo Park, CA, pp. 1-27

[3] Kartardzic, M., 2003, *Data Mining : Concepts, Models, Methods, and Algorithm.* John Wiley and Sons

[4] Kolodner, J., 1993, *Case-based reasoning*. ISBN: 1-55860-237-2, Morgan Kaufmann, San Mateo

[5] Pramudiono, I., 2003, *Pengantar Data Mining : Menambang Pertama Pengetahuan di Gunung Data*, http://www.ilmukomputer.com

Data Warehouse Information Management System RSU Dr. Soetomo for Supporting Decision Making

Silvia Rostianingsih
Department of Informatics Engineering, Faculty of Industrial Technology, Petra Christian University
Siwalankerto 121-131, Surabaya 60236, Indonesia
+62-31-2983455
silvia@petra.ac.id

Oviliani Yenty Yuliana
Department of Informatics Engineering, Faculty of Industrial Technology, Petra Christian University
Siwalankerto 121-131, Surabaya 60236, Indonesia
+62-31-2983455
ovi@petra.ac.id

Gregorius Satia Budhi
Department of Informatics Engineering, Faculty of Industrial Technology, Petra Christian University
Siwalankerto 121-131, Surabaya 60236, Indonesia
+62-31-2983455
greg@petra.ac.id

Denny Irawan
Department of Informatics Engineering, Faculty of Industrial Technology, Petra Christian University
Siwalankerto 121-131, Surabaya 60236 Indonesia

ABSTRACT
Dr Soetomo General Hospital in Surabaya has computerized their administration system base on the Indonesian Health Department standard. It calls Sistem Informasi Rumah Sakit (SIRS) and it only generates structured information. So far, SIRS can not generate unstructured information for supporting decisions making.

Due to the problem, this research focuses on develop Data Warehouse and Online Analytical Process (OLAP) Tools for supporting decision making about inpatient, payment, and surgery. The Application includes a transformation process from SIRS database into OLAP Warehouse database, and processing OLAP Warehouse database into multidimensional pivot table, and generating graphics. The application was developed using Oracle 9i as the database and Net Beans 5.5 as programming language.

Keywords
data warehouse, OLAP, healthcare information, inpatient

1. INTRODUCTION
Soetomo is one of the biggest hospitals in the east Indonesia. Soetomo has computerized their administration system since 2002 using Oracle 9i Database and Oracle Developer. The system is standardized by the Indonesian Health Department and it calls Sistem Informasi Rumah Sakit (SIRS) [1]. SIRS supports periodic reports as structured information that is needed by the health department. The report is produced at monthly basis.

The health department or hospital director often requests inpatient, payment, and surgery unstructured information in several formats (multidimensional) to Soetomo. Unstructured information is not supported by SIRS. In addition, Soetomo does not have Oracle Warehouse Tools for generating the needed unstructured information. Moreover, Soetomo does not have a full time programmer to develop and maintain application system. To full fill the needed information, Soetomo must print several structured reports that relate with the needed information and combine several information use Microsoft Excel to become another report.

Based on the problems, Soetomo need an OLAP tools for generating unstructured report in multidimensional and hierarchal view.

2. MODEL, ANALYSIS, DESIGN, AND IMPLEMENTATION

2.1 Data Warehouse in Health Care
One of the key aspects for a healthcare data warehouse design is to find the right scope for different levels analysis. The analysis of healthcare outcomes is proposed to find scope studies of treatment progress for the next visit. These scopes allow the database to support multi levels analysis, which is imperative for healthcare decision making [2].

The complexity of data analysis determines the number of patients in the risk group for a particular disease. Disease risk levels must be set and adjusted on a regular basis to ensure the coverage of all patients in a care management [3].

2.2 Data Warehouse Methodology
In each state, patient records are registered at a various locations. These records are heterogeneous due to their different source, interpretation, and purpose. There are a number of different stakeholders with the different goals, who have access to the different sources of data. Centralization allows various analysis, data flow, and process mining to search for global estimations and global understanding of a hidden knowledge.

Turning the specific clinical domain information to a Clinical Data Warehouse (CDW) can facilitate efficient storage, enhances timely analysis and increases the quality of real time decision making processes [4].

There are six steps for building a medical data warehouses [5].

- Identify the requirement for the building of the data warehouse

The developers should be able to convince the stakeholders that they have a sound solution for this critical requirement.

- Quality and scope of the sources

Identify the quality and scope of each data source and also the rate of updating (depend on the dynamics of the entities to which the data refers).

- Identify what data is needed by the stakeholders

Match the potentially collectible data with the results that are desired by the stakeholders and the decision makers.

- Build an ontology

In distributed environments, the denominators for data attributes and values can be different.

- How to update the central repository

Establish the update policy for each local source and estimate the costs involved.

- Enact exception handling protocols

The sources should be analyzed with the simplest methods and after the data collected should be analyzed immediately with the same methods to detect anomalies that are induced by the data gathering process.

A star schema is consisted of fact tables and dimension tables. The fact tables describe business fact during a period of time. The dimension tables describe details information for supporting information of fact tables [6].

2.3 System Analysis

The needed SIRS Data to be included into data warehouse are:

- Inpatient room records include roomID and type.
- Room type records include roomType, roomClass, roomRate, and numberOfBed.
- Patient records include patientName, patientAddress, diseaseType, checkInDate, inpatientTime, and roomID.
- Diagnose records include diagnoseType and symptoms.
- Service records include serviceID and serviceName.

OLAP tool is needed for processing warehouse database using pivot table. OLAP tool is built on analysis the purpose [2]. Several unstructured information example often request by the Indonesia Health Department or hospital director are:

- Revenue analysis based on inpatient room, surgery type, room type, and inpatient time. For instance, analyst hospital revenue during a period time.
- Inpatient room utilize based on inpatient room, patient, and time. For instance, analyst the most used room in period time.

2.4 Data Warehouse Star Schema

The conceptual frame work of our research work is shown in Figure 1. There are two main processes, i.e. Create Star Schema and Transformation SIRS OLTP database into OLAP Warehouse database as a preparation and cleansing data and Pivot Table Process as a process to generate multidimensional table and graphic. The first process will discuss in this sub section and the second process will discuss in next section.

Firstly, create or modify the OLAP Warehouse meta schema. A user design or modify a star schema by select tables and fields from the SIRS OLTP meta schema into OLAP Warehouse meta schema as a mapping process. Secondly, transform SIRS OLTP database into OLAP Warehouse along with cleansing data process as shown in Figure 2. Users can transform all data or periodically data.

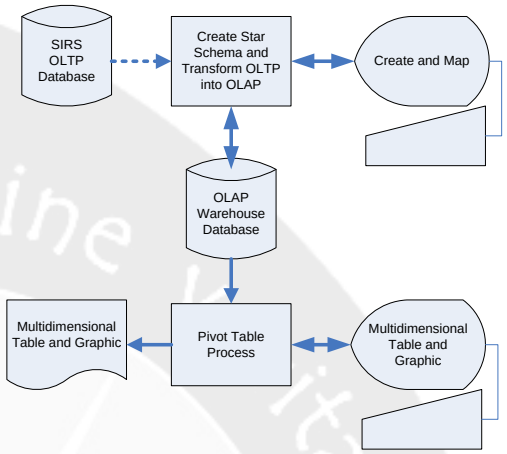


Figure 1. The research conceptual frame

The propose data warehouse star schema consists of three fact tables, i.e. inpatient, payment, and surgery. To make it clear, we describe it into three star schemas. The first star schema is inpatient fact star schema as shown in Figure 3. There are five dimension tables, i.e. time, patient, class, facilities, and room. The second star schema is payment fact star schema as shown in Figure 4. There are three dimension tables, i.e. time, patient, and class. The last star schema is surgical fact star schema as shown in Figure 5. The surgical fact is related with dimension tables time, patient, surgical list, and room.

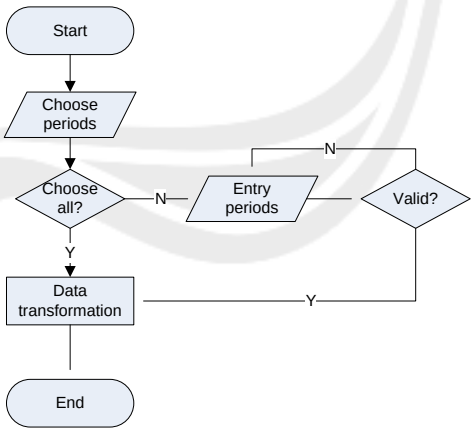


Figure 2. Transformation from SIRS OLTP into OLAP warehouse database

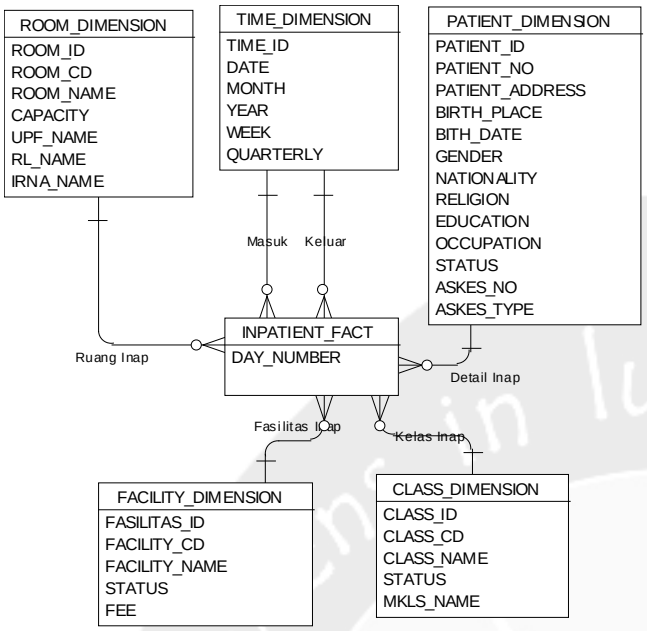


Figure 3. Inpatient fact star schema

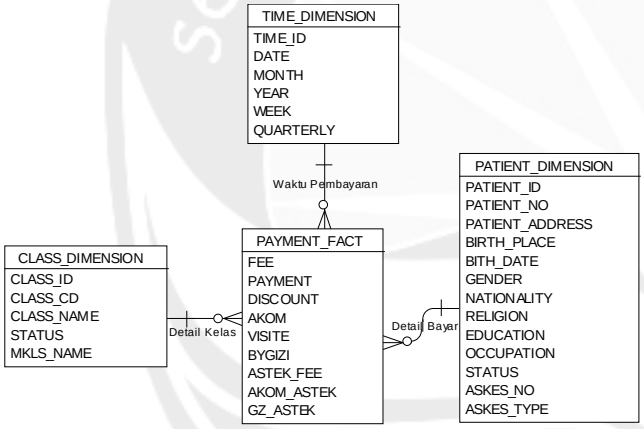


Figure 4. Payment fact star schema

The relation between dimension and fact tables is generally one to many with minimum cardinality optional in fact table. The purpose every dimension table is:

- Time dimension is used for recording the event of inpatient, payment, and surgical. It prepared for multidimensional hierarchical information, e.g. weekly, monthly, quarterly, annually.
- Patient dimension is used for recording the detail patients data who are inpatient and or surgical and pay their medical expenses.
- Facilitates dimension is used for recording the facilities for supporting medical care patient, such as Astek, Askes, Jamsostek.
- Room type dimension to record type of room
- Class type dimension to record class of room
- Surgery lists dimension to record list of surgeries

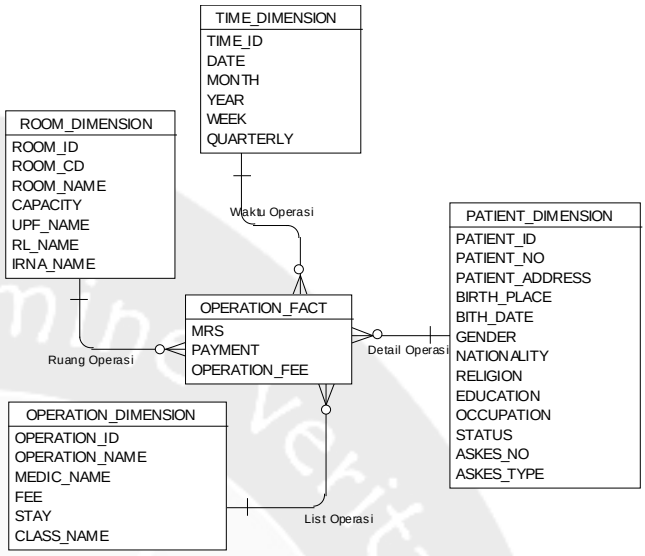


Figure 5. Operation fact star schema

3. THE DESIGN PROCESS

The pivot table process algorithm is shown in Figure 6. The user can set up new parameter or using the saved parameter to process data into multidimensional information as a table or a graphic. The user can select a fact table and several related dimensional tables follow with set attribute to row, column, and data. Furthermore, the user can set the value in the distinct value in the attribute to be included in pivot process. In addition, the user can set the start and end date data to be processed by default is current date. At last, the user can choose the operation process, i.e. count, sum, max, min, and average. The default operation process is count.

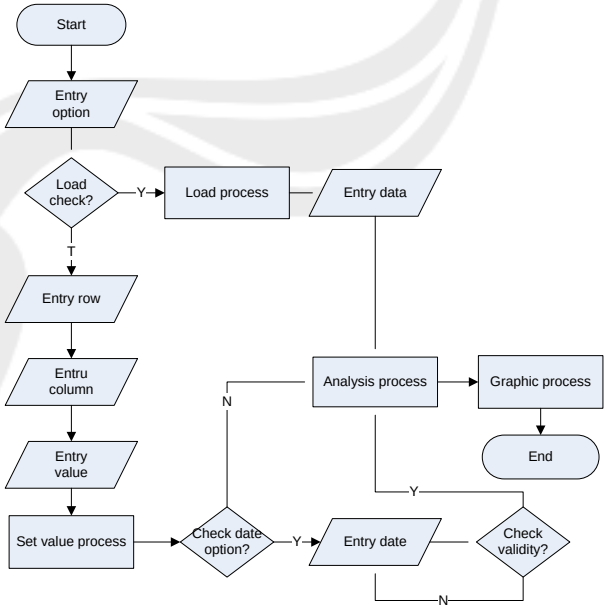


Figure 6. Pivot table process

4. RESULTS

In this section, we demonstrate the developed application system. Pivot table about count of patient based on education vs. room name on February 1–May 1, 2005 is shown in Figure 7. Base on Soetomo requirement analysis, they also want to know the detail information from the pivot table easily. Therefore, we design a pivot popup to zoom out the cell pivot table detail information. For example, the zoom out of *pendidikan Tamat SMTP nama ruang Anak Kelas 1* is shown in Figure 8.

Periode : 02/01/2005 - 05/01/2005

Row : PENDIDIKAN Column : NAMA_RUANG

	Anak Klas 1	Anak Klas 2	Anak Klas 3
Tamat SD	0	0	3
Tamat SMTA	0	0	0
Tamat SMTP	1	0	1
Grand Total	1.0	0.0	4.0

Figure 7. Pivot table count operation in periods

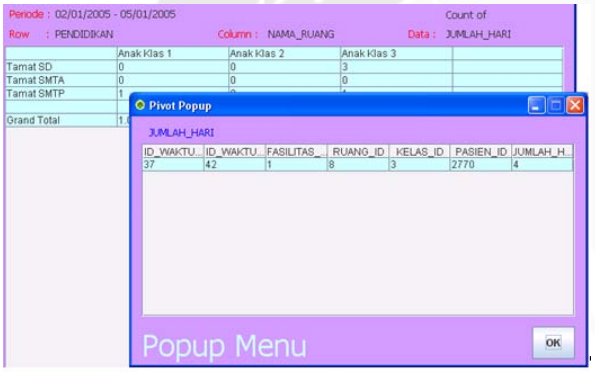


Figure 8. Pivot table zoom out

Figure 9 shows a comparison of Pivot Table Microsoft Excel and output application system for children inpatient based on education and room type in all periods. It shows the same results.

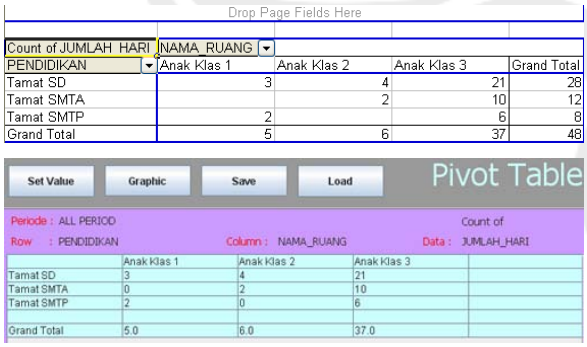


Figure 9. Microsoft Excel vs. output application system comparison

Figure 10 and Figure 11 show a bar chart and a line chart of children inpatient based on education and room type in all periods using count operation.

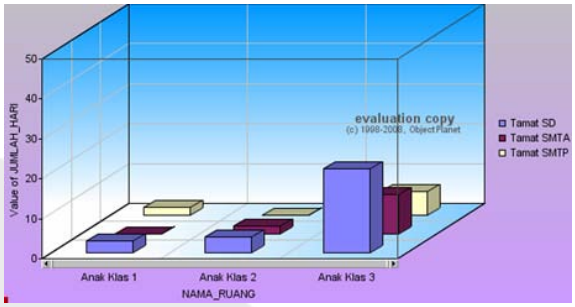


Figure 10. Bar chart education vs. room type in all periods

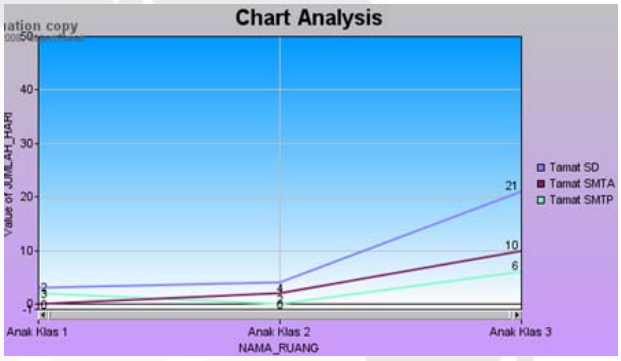


Figure 11. Line chart education vs. room type in all periods

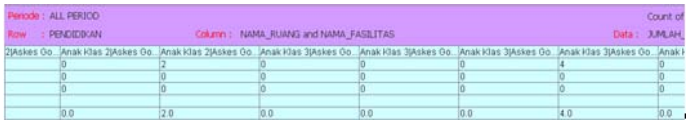
Furthermore we demonstrate the multi dimensional table as the main Soetomo requirement. Figure 12 shows a pivot table with two dimensions on row (education and facilitates name) and one dimension on column (room name). Figure 13 shows pivot table with one dimension on row (education) and two dimensions on column (room name and facilitates name). Figure 14 shows pivot table with two dimension on row (education and facilitates name) and two dimension on column (room name and room class).

Periode : ALL PERIOD Count of

Row : PENDIDIKAN and NAMA_FASI... Column : NAMA_RUANG Data : JUMLAH_HARI

	Anak Klas 1	Anak Klas 2	Anak Klas 3
Tamat SD|Askes Gol I...	0	0	0
Tamat SD|Askes Gol I...	0	0	0
Tamat SD|Askes Gol I...	0	0	0
Tamat SD|Askes Gol I...	0	0	4
Tamat SD|Askes Gol I...	0	0	0
Tamat SD|Askes Gol I...	0	2	0
Tamat SMTA|Askes G...	0	0	0
Tamat SMTA|Askes G...	0	0	0
Tamat SMTA|Askes G...	0	0	0
Tamat SMTA|Askes G...	0	0	0
Tamat SMTA|Askes G...	0	0	0
Tamat SMTA|Askes G...	0	0	0
Tamat SMTP|Askes G...	0	0	0
Tamat SMTP|Askes G...	0	0	0
Tamat SMTP|Askes G...	0	0	0
Tamat SMTP|Askes G...	0	0	0
Tamat SMTP|Askes G...	0	0	0
Grand Total	0.0	2.0	4.0

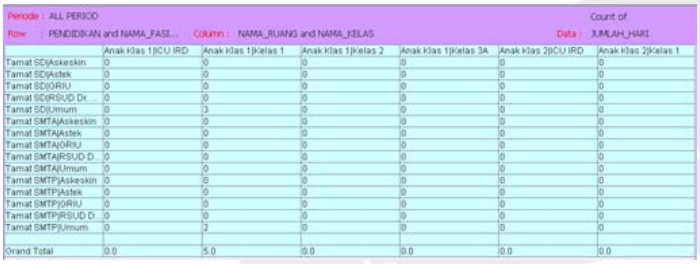
Figure 12. Pivot table with two dimensions on row



Periode : ALL PERIOD — Count of
Row : PENDIDIKAN — Column : NAMA_RUANG and NAMA_FASILITAS — Data : JUMLAH_

2|Askes Go..	Anak Klas 2|Askes Go..	Anak Klas 2|Askes Go..	Anak Klas 3|Askes Go..	Anak Klas 3|Askes Go..	Anak Klas 3|Askes Go..	Anak Klas 3|Askes Go..	Anak K
	0	2	0	0	0	4	0
	0	0	0	0	0	0	0
	0	0	0	0	0	0	0
	0.0	2.0	0.0	0.0	0.0	4.0	0.0

Figure 13. Pivot table with two dimensions on column



Periode : ALL PERIOD — Count of
Row : PENDIDIKAN and NAMA_FASI... — Column : NAMA_RUANG and NAMA_KELAS — Data : JUMLAH_HARI

	Anak Klas 1|ICU IRD	Anak Klas 1|Kelas 1	Anak Klas 1|Kelas 2	Anak Klas 1|Kelas 3A	Anak Klas 2|ICU IRD	Anak Klas 2|Kelas 1
Tamat SD|Askeskin	0	0	0	0	0	0
Tamat SD|Astek	0	0	0	0	0	0
Tamat SD|ORIU	0	0	0	0	0	0
Tamat SD|RSUD Dr...	0	0	0	0	0	0
Tamat SD|Umum	0	3	0	0	0	0
Tamat SMTA|Askeskin	0	0	0	0	0	0
Tamat SMTA|Astek	0	0	0	0	0	0
Tamat SMTA|ORIU	0	0	0	0	0	0
Tamat SMTA|RSUD D...	0	0	0	0	0	0
Tamat SMTA|Umum	0	0	0	0	0	0
Tamat SMTP|Askeskin	0	0	0	0	0	0
Tamat SMTP|Astek	0	0	0	0	0	0
Tamat SMTP|ORIU	0	0	0	0	0	0
Tamat SMTP|RSUD D...	0	0	0	0	0	0
Tamat SMTP|Umum	0	2	0	0	0	0
Grand Total	0.0	5.0	0.0	0.0	0.0	0.0

Figure 14. Pivot table with two dimensions on row and on column

5. CONCLUSION AND DISCUSSION

The developed OLAP tools can be used for generating multidimensional as a pivot table and graphic for inpatient, payment, and surgery. The OLAP tool outcomes are used for supporting decision making and to fulfill the Indonesian Health Department requirements that do not support by SIRS.

This application needed to be improved for run time process, computer memory efficiency, and view of pivot table to be more users friendly.

6. ACKNOWLEDGMENTS

This research is supported by Direktorat Jendral Pendidikan Tinggi, Departemen Pendidikan Nasional (110/SP2H/PP/DP2M/IV/2009) with title "Design and Development of Medical Record Data Warehouse Application System for Supporting RSU Dr. Soetomo Strategic Decisions".

7. REFERENCES

[1] Keputusan Mentri Kesehatan Nomor 1410/Menkes/SK/X/2003. 2003. Sistem Informasi Rumah Sakit di Indonesia (Sistem Pelaporan Rumah Sakit revisi V). Jakarta: Departemen Kesehatan Republik Indonesia.

[2] Parmanto, Bambang. 2005. A Framework for Designing a Healthcare Outcome Data Warehouse. Perspectives in Health Information Management/AHIMA, Volume 2, No 5.

[3] Ramick, Denise C. 2001. Data Warehousing in Disease Management Programs. Journal of Healthcare Information Management, Volume 15, No 2. John Wiley & Sons, Inc.

[4] Sahama, Tony R., Peter R. Croll. A Data Warehouse Architecture for Clinical Data Warehousing. Conferences in Research and Practice in Information Technology, Volume 68. First Australasian Workshop on Health Knowledge Management and Discovery (HKMD 2007).

[5] Szirbik N. B., Pelletier C. 2006 Six methodology steps to build medical data warehouses for research. International Journal of Medical Informatics, Volume 75, Issue 9.

[6] Han Jiawei, Kamber Micheline. 2001 Data Mining: Concepts and Techniques. San Fransisco: Morgan Kaufmann.

Development of An Electronic Medical Record (EMR) In Stayed Nursing Installation

Eko Handoyo
Electrical Department
Diponegoro University
Jl. Prof. Sudharto, SH Tembalang
Semarang, Indonesia 50275
Telephone: +62247460057
eko.handoyo@undip.ac.id

Aghus Sofwan
Electrical Department
College of Engineering
King Saud University
BOX 2454 Riyadh 11451
Kingdom of Saudi Arabia
Telephone: +96614677555
aghus@yahoo.com

Mohammad Muttaqin
Electrical Department
Diponegoro University
Jl. Prof. Sudharto, SH Tembalang
Semarang, Indonesia 50275
Telephone: +62247460057
moh.muttaqin@gmail.com

ABSTRACT
Nowadays, medical caring need more effective and efficient system, in time, personnel and facility using. The fact that medical record still operates in manual papered medical record which appraised unreliable anymore to handle the medical data, issued idea to convert papered medical record to the electronic one, because of its effectiveness and efficiency. The goal of this research is to create the electronic medical record or known as EMR, form the papered medical record in Ananda Hospital Stayed Nursing Installation. This EMR designed by creating forms that recorded medical data during the patient curing process. Then, the data stored and managed digitally. For each medical data in several forms noted, the system will resulting a code that tell a special information. At last, this system produce the ICD (International Statistical Classification of Diseases and Related Health Problems) code stream consist of codes resulted in medical record forms filling. This stream describe the patient conditions development during the curing process. The stored medical data can be represented as digital medical record.

Keywords
Medical record, stayed nursing, Electronic Medical Record, International Statistical Classification of Diseases and Related Health Problems.

1. INTRODUCTION
Medical record is a fundamental directive in medical serving. Its documentation error issued serving error. The slowness in taking a needed medical data can also cause the slowness in medical serving to the patient who must get exact and quick serve.

The medical record used in medical serving nowadays not always able to give the medical data demanded in time. In its operation, medical staff need to be focus in recording an re-accessing. This duty exactly decrease the medical staff work efficiency, whereas in fact they had to prior their serving to the patient health care activity.

The society health condition monitoring by the governmental medical institute force the medical record processing to produce the scheduled report in ICD (International Statistical Classification of Diseases and Related Health Problems) code stream form. The papered medical record can only produce this stream of code by manual codification, which very sensitive to issued human error and slow, especially if there are a large number of patient handled. Another medical record system was demanded to switch such conventional system, integrated in processing patient health care in internal hospital environment and also can give an effective and accurate report to the right side.

2. TERMINOLOGY
2.1 Medical Record
Medical record born near to the medical science, more than thousand years ago. First it was just a medical documentation in many ancient note. Medical record applied in hospital institution introduced first time as patient registration in 1793. It was then developed in 19th century. Medical record start to organized the index of disease and its complement condition in 1862. In 1871, the creation of Disease Main Index Card was instructed for every patient[5].

In the exposition of section 46, verse 1, *UU Praktik Kedokteran*, medical record defined as the document that record the medical note of examination, curing, medical action and another serving had been being given to the patient[13]. In Indonesia, the medical record that had operate since the Dutch colonization repaired after the publication of *SK Menkes RI No.031/Birhup/1972 tentang Perencanaan dan Pemeliharaan Rumah Sakit*. Chapter 1 and section 3 in that regulation told that the duty to perform the medical record in the hospital. The existence of medical record unit structurally demanded by *Permenkes No.134/Menkes/SK/IV/78*. Medical record operation in physician profession forced by the Instruction in *Fatwa IDI (Ikatan Dokter Indonesia) tentang Rekam Medik (SK No.315/PB/A.4/ 88)*. Another regulation to manage the medical record operation was *Peraturan Menteri Kesehatan RI No.749.a/Menkes/per/XII/1989 tentang Rekam Medik* and *SK Dirjen Pelayanan Medik No.78 Tahun 1991*[3].

In medical record manual issued by Indonesian Medical Council informed that medical record classify into two kind of medical record, the conventional papered medical record and the electronic one[13]. Stimulated by the development of demanded medical record nowadays, the ideal characteristics had to fulfilled for medical record development is the system that electronic (computer aided), accessible, secret, secure, accepted by the clinic staff and patient, and integrated with another information type not patient specified[12]. These ideal characteristics were become a

fundamental reason of migration from papered medical record to electronic medical record system that has a good accommodation ability in increasing the performance of medical care serving, especially in managing the medical data.

2.2 International Statistical Classification of Diseases and Related Health Problems

One important thing produced by medical record operation is the ICD code. ICD (International Statistical Classification of Diseases and Related Health Problems) was published by World Health Organization, a system disease classification and several kind of signs, symptoms, anomaly, complain and external disease cause[14]. ICD is codify form the patient medical diagnosis during care process, including first diagnosis (temporal diagnosis), final diagnosis (main diagnosis), complication disease diagnosis and another addition disease diagnosis that suffered by the patient. The latest version of ICD (version 10), based on Central Java Governmental Health Institute releasing, store more than 2500 codes of disease.

The hospital medical care report to the governmental health institute also reported occasionally in stream of code form, including ICD code for disease. This stream named stream of ICD code. A stream of ICD code produce for every patient caring and consist of 19 group of code (4 of them were the ICD code for disease), with 52 digits length. The creation of this stream of code is very complex because the medical record staff has to choose 4 codes from more than 10000 possibilities just to define the ICD code for disease per stayed nursing case.

3. DESIGN

To facilitate system designing, the system first describe into model using UML (Unified Modeling Language) that performed in several kind of modeling diagrams. The use case diagram is a modeling diagram describe relation among actors and use cases[9] in EMR operation procedure. Use case diagram for EMR system shown in Figure 1 below.

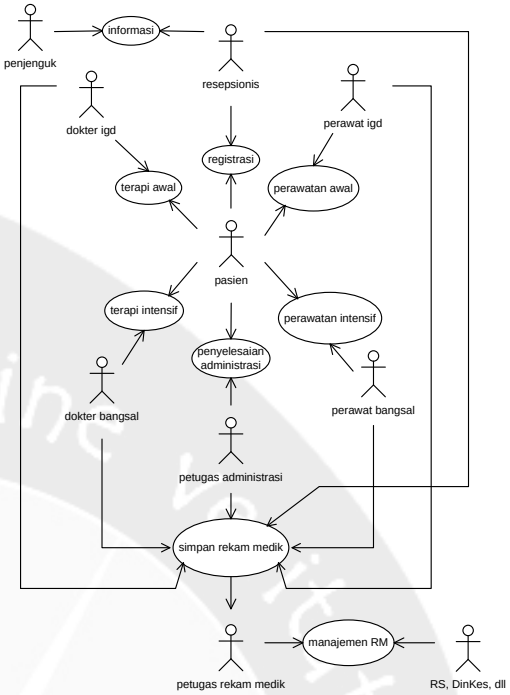


Figure 1. EMR system use case diagram

The activity diagram is the way to simulate all things happen in a use case[9]. Activity diagram designed for every user from all seven user using the system. They are the receptionist, the emergency nurse, the ward nurse, the emergency doctor, the ward doctor, the administration staff and the medical record staff. Medical data accumulated until the end of health care process will be managed by the medical record staff. Figure 2 show all activities of the medical record staff.

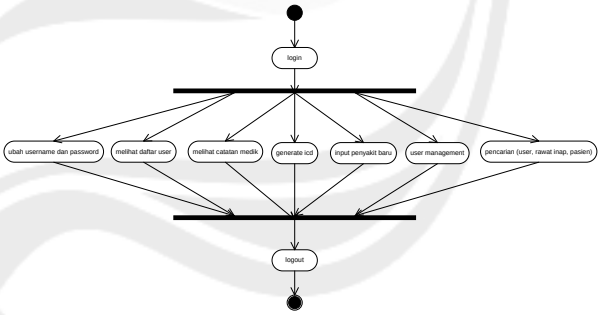


Figure 2. Medical record staff activity diagram

The sequence diagram is a diagram describing the sequence detail of each activity performed in the activity diagram as the order of time or chronologically[9]. An important activity in medical record operation is the ICD stream code generation activity for the patient who had finished his stayed nursing process. This activity done by the medical record staff and can be performed in two different modes. First mode is ICD stream code generation based on the disease which the patient diagnosed. Sequence diagram for this mode is shown in Figure 3.

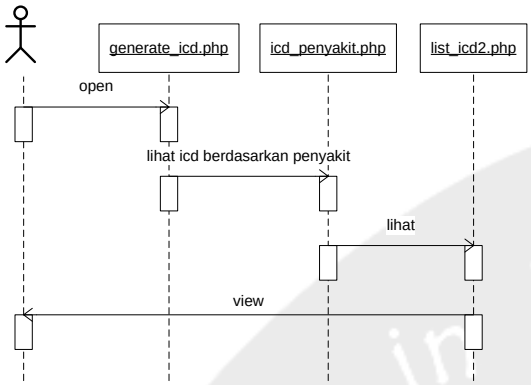


Figure 3. Sequence diagram of ICD stream code generation based on patient disease activity

The rest mode is ICD stream code generation based on the patient guarantor. Sequence diagram for this last mode shown in Figure 4.

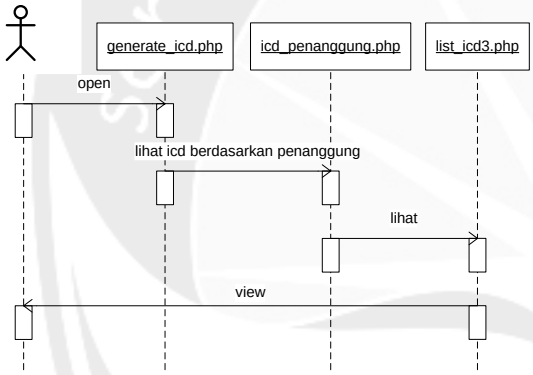


Figure 4. Sequence diagram of ICD stream code generation based on patient guarantor activity

Another modeling diagram is class diagram. This class diagram is a diagram used to perform the classes completed with its packages exist in the development process of the software system[9]. The class diagram also describe the relation among the classes.

4. TESTING

4.1 Unit Testing

Unit testing performed for each unit in the system. Units are commands or menus. The testing executed on login and logout facility, filling and viewing medical note according to the right of each user. The two most important testing is the test to view all the filled medical record and the test to generate the stream of ICD code performed by the medical record staff. Figure 5 show the page to choose which medical note to view, whereas Figure 6 show the testing result of generating ICD stream code based on patient disease diagnosed.

Figure 5. The page show optional medical note to view

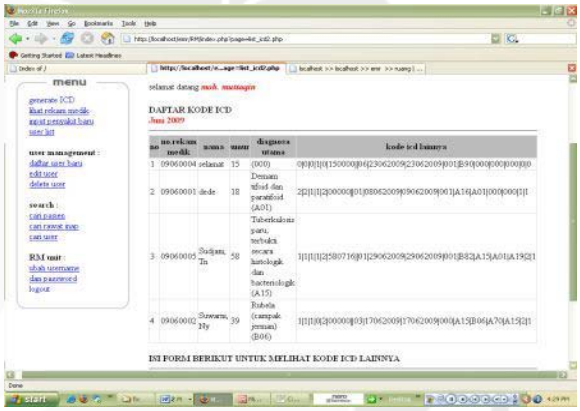


Figure 6. Testing result of ICD stream code generation based on patient disease diagnosed

4.2 System Testing

System testing executed generally to all parts of system and focused on the requirements, use and security aspects. In the requirements aspect, the system enabled to reduce the number operations, automate the filling and avoid its multiplication. Several automatic confirm facilities available to minimize the filling error. Checking facility is also activated in this system to recheck some inputted data so its must be filled with the suitable one. The need to ease accessing, digital recording and quick searching also performed finely. In the use aspect, the medical record functions were fulfilled by the system. Electronic medical record system not only facilitate to write and view the record, moreover can be a large number of active data used for various interest according to the regulation and medical record function development manual. A significant advantage of this system is show on how the ICD stream code can be generated easily from a large number of optional possibilities accurately, in a very short period of time. In the security aspect, the electronic medical record rated better then the conventional kind. The right to write and read managed well refer to the user type, but not broke the rapid interaction among medical staff in the patient health care process. Imitating and counterfeiting action that aliased to another medical staff is hard to execute. To protect the patient, this system completed with the documentation of every medical action noted so

if any malpractice occurred, the proof of the accident will be recorded clearly.

4.3 Acceptance Testing

Acceptance testing show that the system is completed and perform according to the requirements that being the basis of designing process and also accepted by all users operate the system soon, that is the medical staff work in stayed nursing installation. Beta test performing by the real users from various units, consist of registration, emergency, ward, administration and medical record unit, plus general service unit, in stayed nursing installation Ananda General Hospital, Salatiga.

5. CONCLUSION

This stayed nursing EMR system is an integral part of hospital medical record connecting structure. As shown in the research purpose, this stayed nursing EMR system had reached its goal to perform digitalization of the papered medical record use in stayed nursing installation and became reliable system in accessing speed, accurate in reporting and improve the work efficiency of medical staff who use it. This stayed nursing EMR system belong to 7 users (receptionist, emergency nurse, emergency doctor, ward nurse, ward doctor, administration staff and medical record staff) from 5 medical care units in stayed nursing installation (registration unit, emergency unit, ward, administration unit and medical record unit). Output of the system can be reused to be the raw data for creating another medical document note, especially ICD. This stayed nursing installation applied the latest version of ICD code, ICD code version 10, released by Central Java Governmental Health Institute.

6. REFERENCES

[1] Ginsburg, Mark., "Pediatric Electronic Health Record Interface Design: The PedOne System", in Proceedings of The 40th Hawaii International Conference on System Sciences, IEEE 2007, pp. 1530-1605.

[2] Hakim, Lukmanul., Membongkar Trik Rahasia Para Master PHP, Cet. I. Lokomedia, Yogyakarta, 2008.

[3] Hanafiah, M. Jusuf dan Amri Amir., Etika Kedokteran dan Hukum Kesehatan, Ed.3, Penerbit Buku Kedokteran EGC.

[4] Harlan, Johan., Dari Rekam-Medik Kertas ke Rekam-Medik Elektronik, Presentasi, Pusat Studi Informatika Kedokteran, Universitas Gunadarma, Jakarta, without published year.

[5] Infokez., Sejarah dan Perkembangan ilmu Rekam Medis, http://www.infokez.wordpress.com / Sejarah & Perkembangan ilmu Rekam Medis « Line Of Infokez.htm, Juni 2008

[6] J. Kairouz, A. Lam, A.S. Malowany, F.A. Carnevale, R.D. Gottesman., "A Vital Sign Monitoring System for a Pediatric Intensive Care Unit", Seventh Annual IEEE Symposium on Computer-Based Medical Systems, SM4: Signal Processing 3, 1994, pp. 217-222.

[7] Kadir, Abdul., Penuntun Praktis Belajar SQL, Ed. 1. Andy Offset, Yogyakarta, 2002.

[8] Muttaqin, Moh and Eko Handoyo., "Designing A Web Base Electronic Medical Record in Stayed Nursing Installation", International Graduate Conference on Engineering and Sciences, D35: Computing and Information Technology, 2008.

[9] Nugroho, Adi., Rational Rose untuk Pemodelan Berorientasi Objek, Cet.I.17, 39-40, 51-52, 61, 92, 110. Informatika, Bandung, 2005.

[10] Nugroho, Bunafit., Database Relasional dengan MySQL, Ed.1.1-4. Andy Offset, Yogyakarta, 2005.

[11] Nugroho, Bunafit., Trik dan Rahasia Membut Aplikasi Web dengan PHP, Cet. I, Ed.1.47. Gava Media, Yogyakarta, 2007.

[12] Sanjoyo, Raden., Aspek Hukum Rekam Medis, http://www.yoyoke.web.ugm.ac.id, D3 Rekam Medis FMIPA Universitas Gadjah Mada

[13] Rusli, Arsil, dkk., Manual Rekam Medis, Konsil Kedokteran Indonesia/Indonesian Medical council, Jakarta, 2006

[14] ---, ICD, http://id.wikipedia.org/wiki/icd.htm

Development of Supporting Sales Analysis Application Using Frequent Closed Constraint Gradient Mining Algorithm (FCCGM)

Susana Limanto
Informatics Engineering
University of Surabaya
Raya Kalirungkut, Surabaya 60292 Indonesia
(031)2981395
susana@ubaya.ac.id

Dhiani Tresna Absari
Informatics Engineering
University of Surabaya
Raya Kalirungkut, Surabaya 60292 Indonesia
(031)2981395
dhiani@ubaya.ac.id

ABSTRACT

Making a promotion on certain items is a strategy to gain the competitive advantage in retail business. If a customer purchases item A for example, promotion on item A can be done by giving a bonus of item B, providing discount for item B, or any other form of promotions. At first glance, each kinds of promotion may decrease or even make harm the retail business. To avoid these conditions, the retail entrepreneur must be able to obtain proper information by mining their sales data. Information that can be gained is about set of items were purchased usually by customers in conjunction with particular items which appears to increase the average profit of that particular item. Based on this information, the sales volume of any particular item may be raised by making the particular item as a promotion item that must be sold together with certain set of items.

The most common obstacles in getting information from sales data is the limited human ability to process large amounts of data efficiently. This research is proposed to help entrepreneurs to simulate the retail sales data in generating information about the set of items that frequently purchased with a particular item which will increase the average profit of a particular item, using the FCCGM algorithm.

Keywords

data mining, frequent closed constrained gradient mining, sales analysis.

1. INTRODUCTION

Recently, retail business has been became a business that attract many entrepreneur attentions. This can be proved by the increasing number of existing retail businesses at the nearby locations, such as Alfa Mart, Alfa Midi, Indomaret and Circle K. The number of emerging retail business will increase business competition. So, many promotions are made to attract as many customers to come to their stores.

Good promotion should not harm the respective retail businesses and could attract customers. Unfortunately, not all promotion attracts customers. Retail entrepreneur must be able to establish a strategy based on the customer's shopping habits to avoid condition as mentioned before. Information about customer's shopping habits can be gained from sales data and can be used to get information about the set of items that were usually purchased in conjunction with certain items. The set of items that are sold together with some promotion items might increase or decrease the average profit of those promotion items. *Frequent Closed Constrained Gradient Mining* (FCCGM) is an algorithm which can be used to obtain information about the set of items that will increase the average profit of promotional items.

2. FREQUENT CLOSED CONSTRAINED GRADIENT MINING (FCCGM) ALGORITHM

FCCGM algorithm can be used to determine set of items that were usually purchased by customers in conjunction with any promotional item, which may increase average profits of promotion item as desired. Those set of items is known as frequent closed constrained gradient item sets. This algorithm was first introduced by Wang et al. in 2006 and was implemented on sales data (also known as transactional data) in a retail database. This algorithm comprises six main processes. The processes are calculation of measure value, followed by construction of projected database, deletion of the items that do not meet *the minimum* criteria for *support* and *gradient threshold*, construction of FP-Tree, deletion of items that do not satisfy the gradient threshold and the last process is to mine remaining data. All of these processes will be explained below.

The first process is calculating the *measure* value. *Measure* is defined as average profit of promotion items in each transaction. After the calculation of measure value, projected database is constructed. Projected database is constructed by excluding the transactions without promotion items, removing the promotion items from remaining transactions, and then sorting remaining items in all transactions according to support value in descending order. Support value is defined as the frequency of appearance a group of specific items in all existing transactions **(Megaputer, 2000).**

The third process consists of few steps. The first step is to remove the items that do not meet the *minimum support* from *the projected database* and to calculate *the gradient threshold* value. *Gradient threshold* value is defined as sum of all transaction *measure value* in *the projected database*, divide by the number of transactions and time by defined *minimum gradient*. The next step is to calculate the top-*K average value* of each item. *Top-K average* value of an item Z is average *measure* of the first K

transactions containing items Z that has the largest *measure* to satisfy the condition that K is not smaller than *the minimum support* **(Lam, 2007).** Items with average *top-K value* less than the *gradient threshold* will be removed.

In the fourth process, FP-Tree was constructed based on the remaining items from the previous process. *FP-Tree* is a tree data structure that is used to store transactional data. *FP-Tree* is constructed by mapping each transaction data into a path in it. *FP-Tree* data structure will works more effective if a retail database consists of many data transactions with the same items. Those data transactions will be recorded at the same path. Each *node* in *FP-tree* contains few information. They are item code, support value and *measure value*. The *support* and measure value will increase every time the *node is passed*.

After the *FP tree* was constructed, the next *process would be* the calculation of the *Top-X Average* value for each item / node in the *FP Tree*. *Top-X average* of item Z is an average *measure* of the Top-X *nodes* of the item Z where X is the smallest number satisfying that the sum of top-X *node* count is no smaller than *minimum support* **(Wang, Han, Pei, 2006).** The items with *Top-X average* value less than *the gradient threshold* will be pruned from the *FP tree*. The frequent closed item set mining as the sixth process, will be implemented on remaining items from the fifth process. LCLOSET algorithm is used to perform mining process. This algorithm is a part of the CLOSET+ which is specifically designed to mine *frequent closed item set* on the sparse database. Sparse database is the characteristic of retail database **(Wang, Han, Pei, 2003).** The items generated from this sixth process are the frequent closed constrained gradient itemsets of promotional items that can increase the average profit of promotion items.

3. DISCUSSION AND RESEARCH RESULT

This study began by analyzing few retail businesses to determine the system requirements. The analysis results against several number of retail businesses showed that:

1. Most entrepreneurs usually consider sales data has no significant meaning, so those data are rarely specifically analyzed. They prefer to rely on their feelings and experiences to make a *marketing* strategy plan.

2. Commonly, many retail database are stored in different media, such as Microsoft Excel, Microsoft Access, Oracle or manually stored. It may cause the difference in data name convention. For example, data items, item master, good master are actually refers to the same thing. The difference of media storage and data name convention may initiate some difficulties in implementing FCCGM algorithm **(Wijaya, 2009).**

Beside the analysis of several retail businesses, an analysis for FCCGM algorithm was also performed to determine the desired inputs and outputs needed by this algorithm. Input data required by this algorithm comprises master of item data, master of item group data, sales/transactional data, promotion item, minimum support threshold, and the percentage of desired average profit. Item master data at least, should contain information about item code, item name, selling price average and item group code. Master of item group data should contains information about the group code and group name. Sales/transactional data should include information about the sales transaction number, quantity and selling price of every sold item in each transaction. Any additional information may occur in each master table as needed. **(Wijaya, 2009).** The outputs generated by the FCCGM algorithm are item sets that would increase the average profits of defined promotion items that satisfy a certain gradient threshold, when they were sold together, which is usually called as *frequent closed constrained gradient item sets.*

After the analysis had accomplished, design step was performed based on the analysis results. There are three kinds of design that must be completed. They are data design, process design and *user interface* design. Data design comprises three different tables and four variables, namely:

1. Setting Table. This table is used to record the equivalent of the table name and information name between current retail business tables and tables are used in application programs. This Setting Table is used to overcome the differences of data name convention between retail database and database in the developed application programs. Two main information that must be exist in Setting Table are namaRetail that contains list of data name that is used in current retail database and namaProgram which contains the data name that is used in application program.

2. Master Item Table. This table contains information such as Item Code, Item Name, average of buying price and item group code.

3. Master Item Group that contains information about group code and group name.

4. Sales Data Table that contains information about transaction number (tid), code, quantity and selling price of every sold item in each transactions.

5. MinSupport variable records information about single item appearance frequency that is allowed in all transactions data to be used in subsequent processes. Minimum support value must be greater then zero.

6. MinGradient variable states the percentage of desired average profit increasing of promotional items.

7. DtSetProbeItem variable records the list of promotional items.

Process design will be describe in using flow chart diagram in **Figure 1** below.

The difference of data storage media between the retail business database may cause difficulties in application development process. To overcome this problem, database format used in this application is in Microsoft Access format only. The overall application process design could be seen in Figure 1. The setting process used to record the equivalent of the table name and the information name between the table in retail database and the table in application program. Result data from this process will be used as data in import

process. Import process transforms data that was taken from retail database into the corresponding database in applications program. FCCGM process starts after the data had imported.

To give additional benefit for the entrepreneur in planning a marketing strategy, the output of the application is designed not only to display a list of item sets that can increase the average profit of promotion items *(frequent closed item set.)*. Additional features were also provided by the application are:

1. Promotion items can be selected based on items list or item groups list.
2. Frequent closed item sets can be displayed in item list or items groups list.
3. Frequent closed item sets history.
4. Capability to sort the displayed data either *ascending* or *descending* order.
5. Capability to sort the displayed data by the item name, by *support value* or by additional profit.

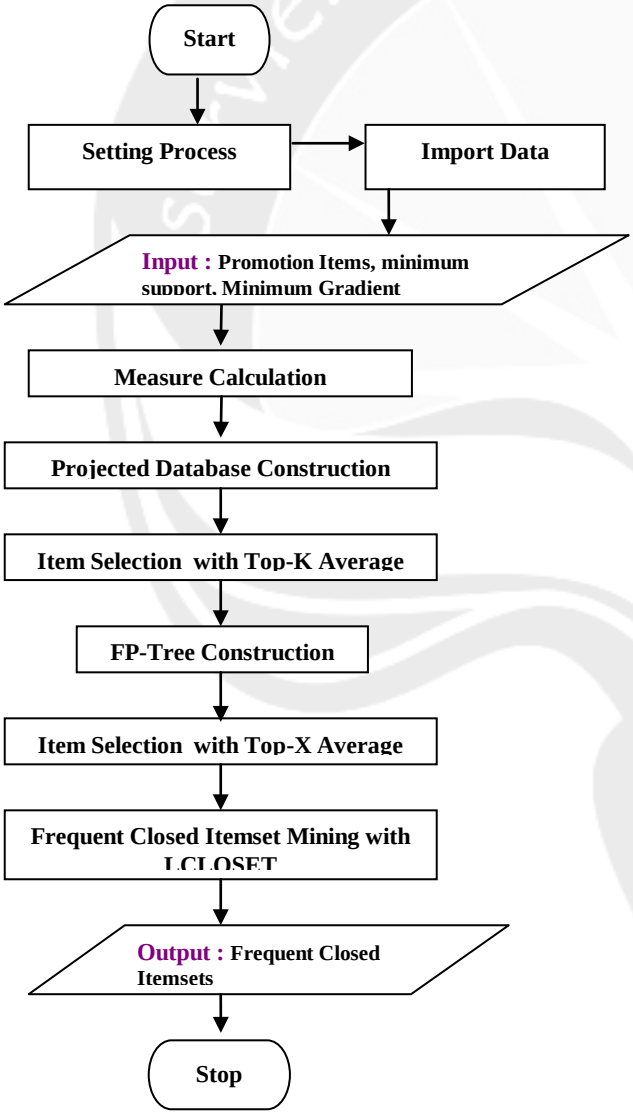


Figure 1. Flowchart of process design

Design result was implemented using Visual Basic.NET 2003 and Microsoft Access as its database management system. Implementation example can be seen in **Figure 2.** Three testing were performed to ensure that the application has been running in accordance with the desired requirements. The first time, we perform a test to ensure that FCCGM already running correctly on the application program. Testing is performed by comparing the manual calculation results with the results of the application program. Data shown in Table 1 to Table 3 are data used for this testing. After several improvements can be ascertained that FCCGM already running correctly on the application program. Next we perform parameters testing. Objective of this parameter testing is to discover how any parameters may influence process speed. Parameter testing was performed by giving five variations of sales transactions dataset that was combined by three variations of gradient threshold. Five variations of sales transactions dataset were used are contain 2500, 5000, 7500, 10,000, and 12,500 transactions. The three variations of gradient threshold value were used are 0%, 5%, and 10%. Minimum support value was used in all testing are 5%. The last testing is performed to ensure that the application program can really help businessman to develop sales strategy. The test is performed by demonstrating the application program to a businessman. The businessman stated that the application can help marketing program but with some additional features, such as the history facility. After all suggestions from the businessman are met, the businessman stated that the application program is good enough to help in developing some marketing strategy.

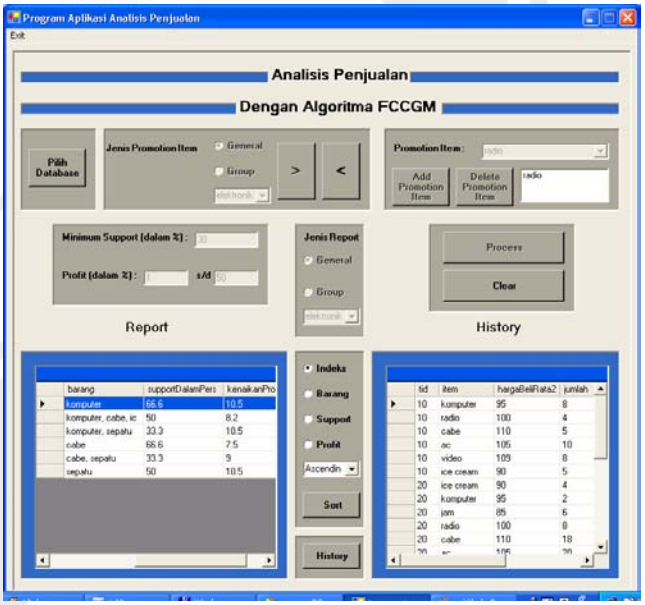


Figure 2. Example of implementation result

The test results show that application output is in conformity with the results of manual calculations, and all facilities are running as it should. The parameter test results are presented in Table 4 to Table 6 (Wijaya, 2009).

Table 1. Table master items

Item Code	Name	Average of buying price	Group Code
a	ice cream	150	J2
b	sepatu	120	J3
c	komputer	130	J1
d	Jam	140	J3
e	Radio	100	J1
f	telepon	110	J2
g	kulkas	125	J1
I	gunting	135	J3
k	Vcd	145	J1
l	Juice	123	J2
m	Ac	105	J1
n	koran	115	J4
p	video	109	J1

Table 2. Table group of items

Group Code	Group Name
J1	elektronik
J2	makanan
J3	peralatan
J4	lain-lain

Table 4 to Table 6 show that the growing number of sales data will increase processing time on average, except in sales data consists of 10,000 transactions. The observation indicates that this variance happens because there are many items do not meet minimum support and gradient threshold value. It causes processing time shorter because of the small amount of data were processed. The testing results also indicate that the increasing number of desired gradient threshold value will cause the decreasing of processing time. It happens because the increasing of gradient threshold value will decrease the number of item sets which meet the criteria. The decreasing number of items will finally reduces processing time

Tabel 3. Table sales

tid	Item Code	Quantity	selling price of every sold item
10	A	5	175
10	C	8	160
10	E	4	125
10	F	5	200
10	M	10	210
10	p	8	195
20	a	4	180
20	c	2	175
20	d	6	165
20	e	8	139
20	f	18	190
20	m	20	235
20	p	4	170
30	a	9	160
30	b	5	179
30	c	2	180
30	e	6	150
30	f	4	190
30	g	7	200
30	m	8	210
40	b	3	185
40	e	9	126
40	f	7	180
40	I	5	195
50	b	6	190
50	c	9	173
50	e	4	145
50	n	7	167
50	p	5	183
60	k	3	170
60	l	2	183

Table 4. Parameter Testing Result with Gradient Threshold = 0%

Number of Transactions	2500	5000	7500	10000	12500
Promotion Item	Nokia	Nokia	Nokia	Nokia	Nokia
Minimum Support	5	5	5	5	5
Profit Value	0%	0%	0%	0%	0%
Time (in seconds)	57	552	1576	992	3821

Table 5. Parameter Testing Result with Gradient Threshold = 5%

Number of Transactions	2500	5000	7500	10000	12500
Promotion Item	Nokia	Nokia	Nokia	Nokia	Nokia
Minimum Support	5	5	5	5	5
Profit Value	5%	5%	5%	5%	5%
Time (in seconds)	47	365	1105	795	2936

Table 6. Parameter Testing Result with Gradient Threshold = 10%

Number of Transactions	2500	5000	7500	10000	12500
Promotion Item	Nokia	Nokia	Nokia	Nokia	Nokia
Minimum Support	5	5	5	5	5
Profit Value	10%	10%	10%	10%	10%
Time (in seconds)	43	261	870	671	2362

4. CONCLUSION

Application program created with some various facilities can help the businessman to develop marketing strategies. Given facilities allow businessman to analyze every item that become candidates to increase profit. Moreover, from the test results some conclusion could be taken. First, the test results showed that the process is influenced by many things, like minimum support value, the amount of sales data, the average profit value and the level of sales data variation. In general, the greater number of sales data, the longer time takes to run FCCGM algorithm. The greater minimum support value, sales data variation and gradient threshold value, the more data exit the criteria. It will cause the decreasing number of the data used in the process and will make the time required to run the algorithm FCCGM faster.

5. REFERENCES

[1] Lam, J., 1997, "Multi Dimensional Constraint Gradient Mining", Thesis, Simon Fraser University, Canada

[2] Megaputer Intelligence Inc., 2000, *Market Basket Analysis*, Moscow, Megaputer Intelligence Inc., diambil dari: http://www.megaputer.com/tech/wp/mba.php3 [Diakses tanggal 10 September 2004].

[3] Wang, J., Han, J., and Pei, J., 2006, Closed Constrained Gradient Mining in Retail Database, *IEEE Transaction on Knowledge and Data Engineering*, Vol 18, No.6, page 764-769.

[4] Wang, J., Han, J., and Pei, J. 2003. "CLOSET+: Searching for the Best Strategies for Mining Frequent Closed Itemsets", http://www-faculty.cs.uiuc.edu/~hanj/pdf/kdd03_closet+.pdf

Tanggal akses : 26 Agustus 2008

[5] Wijaya, H., 2009, Pembuatan Program Aplikasi Analisis Penjualan Barang Berbasis Algoritma Frequent Closed Constrained Gradient Mining, Universitas Surabaya.

Implementation of KMS to Integrate Knowledge Management and Supply Chain Management Process

Vivine Nurcahyawati	Retno Aulia Vinarti	Mudjahidin
Sekolah Tinggi Manajemen Informatika & Teknik Komputer Surabaya	Institut Teknologi Sepuluh Nopember Surabaya	Institut Teknologi Sepuluh Nopember Surabaya
Program Studi Sistem Informasi	Gedung Teknologi Informasi – ITS	Gedung Teknologi Informasi – ITS
Jl. Raya Kedungbaruk 98, Surabaya 60298	Kampus ITS, Sukolilo, Surabaya 60111	Kampus ITS, Sukolilo, Surabaya 60111
Telp. 031 8721731	Telp. 031 5922949	Telp. 031 5922949
Fax. 031 8710218	Fax. 031 5964965	Fax. 031 5964965
vivine@stikom.edu	zahra_17@is.its.ac.id	mudjahidin@its-sby.edu

ABSTRACT

This paper aims to build a Knowledge Management System, which is an integration of KM with SCM, this KMS will be used by Manufacturer, Distributor, Wholesaler and Retailer with the KMS.

KMS will integrate data, information and knowledge from all part of Supply Chain, so hope this KMS can help reduce inventory costs, reduce ordering costs, increase sales, and will reduce the sales price to the customer's hand.

Keywords

Supply Chain Management, Knowledge Management

1. INTRODUCTION

Based on economic principle, the fundamental principle of economic activity, this principle reveals itself in man's endeavour always and everywhere to attain the highest possible satisfaction with the least possible sacrifice (labour, money) in every task, production, distribution or consumption [3], so the companies compete to find the cheapest price for best quality.

This paper proposes to create a KMS that will integrate information and knowledge within SCM process from Manufacturer to Wholesaler that allows for e-fulfillment and e-procurement for all members of the supply chain, which is expected to reduce the final price to be received by the customer with the best possible use of Promotions are usually issued periodically by the Manufacturer.

Besides the cheaper price, KMS will also reduce bullwhip effect, the causes of bullwhip effect are because customer demand is rarely perfectly stable, all part of Supply Chain must forecast demand to properly position inventory and other resources. Forecasts are based on statistics, and they are rarely perfectly accurate. Because forecast errors are a given, companies often carry an inventory buffer called safety stock. Moving up the supply chain from end-consumer to raw materials supplier, each supply chain participant has greater observed variation in demand and thus greater need for safety stock. In periods of rising demand, down-stream participants increase orders. In periods of falling demand, orders fall or stop to reduce inventory. The effect is that variations are amplified as one move upstream in the supply chain (further from the customer).

The causes can further be divided into behavioral and operational causes:

Behavioural causes

- misuse of base-stock policies
- misperceptions of feedback and time delays
- panic ordering reactions after unmet demand
- perceived risk of other players' bounded rationality

Operational causes

- dependent demand processing (forecast Errors and adjustment of inventory control parameters with each demand observation
- Lead Time Variability (forecast error during replenishment lead time)
- lot-sizing/order synchronization (consolidation of demands, transaction motive, quantity discount)
- trade promotion and forward buying
- anticipation of shortages (allocation rule of suppliers, shortage gaming, Lean and JIT style management of inventories and a chase production strategy)

From Figure 1 we can see perfectly the different order quantity from End Customer, Retailer's order to Wholesaler, Wholesaler's order to manufacturer and Manufacturer's Order to Supplier are increases and more unstable in time range.

Figure 1. The bullwhip effect in action

2. LITERATURE

2.1 Supply Chain Management

A supply chain consist of all parties involved, directly or indirectly in fulfilling a customer request, the supply chain not only includes the manufacturer and suppliers, but also transporters, warehouses, retailers and customer themselves, the supply chain also includes all function but are not limited to new product development, marketing, operations, distribution, finance and customer services [2].

Supply Chain strategy that used in this paper is CPFR (Collaborative Planning, Forecasting and Replenishment), this strategy is a Collaboration initiative between retailers and their supplies, based on concept that sharing demand information between tiers in the supply chain improves overall performance in terms of in-stocks rates, inventory and sales [6].

In this paper, we suggest that all part of Supply Chain do Collaborative Planning and then Forecasting demand together. CPFR and other collaborative processes provide true end-user benefits. Supply chain collaboration is more than visibility, information sharing and improved technology. It also involves changing the nature of trading relationships in order to add value for end users, as well as benefiting all participants in a collaborative value chain [8].

The purpose of CPFR is to improve partnerships and to facilitate all participant of Supply Chain responsiveness through collaborative processes and information sharing, the steps are collaborative arrangement, sales forecast, order forecast and then generate order between seller and buyer, can be seen on Figure 2.

After doing CPFR, then all part of Supply Chain must commit to use KMS to share their knowledge and information, in order to achieve the final goal, that is make the cheapest way to reach the best product, that will take many customers.

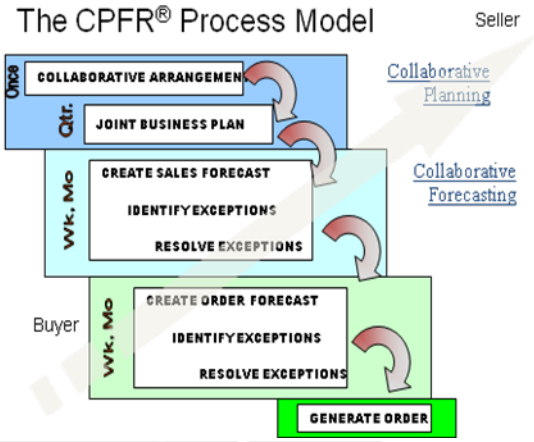


Figure 2. The collaborative planning, forecasting and replenishment

2.2 Knowledge Management

KMS should not be seen as a voluminous centralized data base. They can rather be imagined as large networked collections of contextualized data and documents linked to directories of people and skills and provide intelligence to analyze these documents, links, employees' interests and behavior as well as advanced functions for knowledge sharing and collaboration.

Goals of using KMS are for example to generate, share and apply knowledge, to locate experts and networks, to actively participate in networks and communities, to create and exchange knowledge in these networks, to augment the employees' ability to learn and to understand relationships between knowledge, people and processes [5].

There are four modes of knowledge conversion [4]. They are:

1. Socialization, from tacit knowledge to tacit knowledge
2. Externalization, from tacit knowledge to explicit knowledge
3. Combination, from explicit knowledge to explicit knowledge
4. Internalization, from explicit knowledge to tacit knowledge

To integrate Knowledge Management with Supply Chain Management, we need data storage and data warehouse to save all knowledge and information that stored by KMS in every user. The information that would be shared are inventory stock, sales forecasting and promotion system of Manufacturer.

According to [7], to make externalization of tacit knowledge is not easy, tacit knowledge is difficult for organizations to exploit. Since it only resides inside people, it cannot be easily be sought electronically like KMS. The problem of determining who knows what grows with the size of the organization [1].

3. PROBLEM DESCRIPTION

The organization that suits with this integration is manufacturer that may have a yearly promotion program to be used to achieve final goal for this implementation. Final goal for

this implementation is achieve higher less purchase price of item, in order to make cheaper price to end customer.

The characteristic or model of supply chain participant that can use this KMS are consist of three or four participant, they are Manufacturer that have a promotional program, Distributors, Wholesaler and Retailer.

Usually, manufacturer have a promotions program, a yearly promotions program, the type divided into two based, based on quantity order discount and based on time discount.

As we know, the quantity order discount will be give if the downstream order minimum requirement of promotions, there is two schemes of quantity order discount, the straightforward and complex. The straightforward one usually request one kind of item only, while the complex one request more than one kind of item. For example, the straightforward promotion would be a 10% discount for an order valued between $3,000 and $6,499 for air item A, if orders valued greater than $6,500 would receive a 20% discount. And a complex discount scheme would be a 23% discount for ordering any mixture of item A and item B that totals 30 units.

The quantity order discount usually concern about order value and order quantity, but this type of promotions will make a high bullwhip effect to the manufacturer, because downstream od Supply Chain will focus on this promotion, rather than minimizing their inventory stock in order minimizing inventory holding cost.

Mostly, customer satisfaction is measured by CSI, Customer Satisfaction Index, net profit and minimizing inventory in order to minimize inventory holding stock.

3.1 Knowledge in the Ordering Problem

From a decision theoretic viewpoint, this problem involves

1. Deciding how much to order and when, considering both cost and customer service. There are several facets of the ordering decision that require the application of knowledge. When deciding how many parts to order at any one time, the retailer manager must do these
   a. Determine optimum order quantity for each part, while the UCS generate a recommended order quantity, it is not consider things such as promotional schemes, seasonal trends or product life cycle trends.
   b. Forecast of sales that is estimated based on look back over time that computes the historic average sales volume per day.
2. The error and the uncertainty associated with forecasts are also important in this problem. In practice, uncertainty in sales is addressed using safety stock, based on a desired service level and the expected sales rate.
3. The efficacy of the service parts department incentive structure is also an important part of the knowledge management problem. The combination of CSI, net profit and inventory are conflicting problem, however.

   For example, if a product is in for service and the part is not available at the inventory because of an inventory minimizing goal – which is minimizing inventory holding cost – the product will be shipped from another retailer or wholesaler or maybe from its distribution center incurring higher freight costs for expedited delivery.
4. A time based manufacturer promotions is more efficient than order quantity manufacturer promotions in order to reduce the bullwhip effect on manufacturer. And the promotions usually printed, so it can't be accessed electronically.
5. Data for making the decision are scattered, in Manufacturer, Distributor, Wholesaler, Retailer and the end customer itself.

3.2 A Knowledge Management Approach to a Solution

A knowledge management system (KMS) to assist the manager in his decision-making would be comprised of three components. These are

1. The ability to access promotions electronically, all discount scheme offers should be displayable. Furthermore, the manager should be able to array data these data in any format desired – by expiration date of the offer, by inventory category, by dollar value or some other arrangement.
2. A what-if simulation facility, the simulation engine should provide a recommended solution, as well as allow the manager the capability of running 'what-if' tests before placing an order with upstream.
3. An ability to integrate OEM promotions with the existing online ordering facility, the new system would have the capability of automatically flagging those parts that qualify for a discount and display the quantity required to qualify for discount scheme.
4. Because of CPFR, so the reorder point – the safety stock – must be appropriate with the trend of sales forecasting.
5. All participant of Supply Chain must commit to use KMS and upload their inventory level regularly.

The decision rules for deciding what and how much to based on a combination of historical data that shows sales trends and a desire to order just enough to get discount.

3.3 Business Processes before implementing Knowledge Management System

Here there are the processes that all participant of Supply Chain should do before integration between Supply Chain Management and Knowledge Management to be a Knowledge Management System

1. Each day the Logistic System used by all participant of Supply Chain generates a list of stock numbers with inventory levels that have fallen below the recommended reorder point.
2. Logistic officer then manually scans each order for goods that may qualify for quantity discounts in Manufacturer Promotions, and often adjust the quantities accordingly in order to get specific discount item.
3. The manager checks one to one the availability of order goods in its upstream.
4. And placed the order appropriate, according to their availability.

Picture 2 show the work flow diagram of business processes before implementing Knowledge Management System

But web-based order doesn't always work well because some dealers do not update their records online regularly and not willing to share their inventories with a competitor.

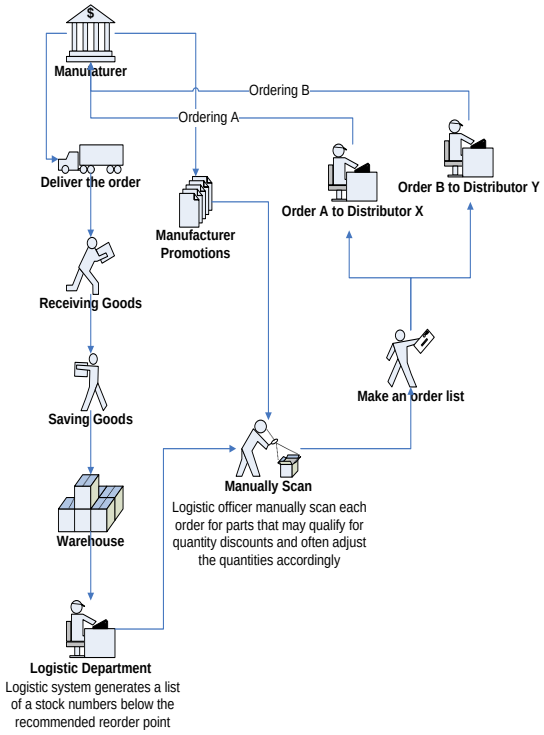


Figure 3. Work flow before implementing KMS

3.4 Business Processes after implementing Knowledge Management System

There is a significant change after the implementation of Knowledge Management System, can be seen on figure 4

1. Logistics officers do not need to manually scan the items that are on the list Logistic System to formulate which will get discount, because there will be a flag.
2. Logistics officers do not need to adjust quantities of items to match the discount to which, because KMS will provide the recommended quantity to order
3. Logistics Officer can also simulate what if by weighing the desired parameters, such as expiry date, dollar amount or category of inventory in order to get the biggest discount if more than one discount is recommended by KMS.
4. Logistics officer can calculate and estimate the reorder point according to the forecast results from the sales history data per month.

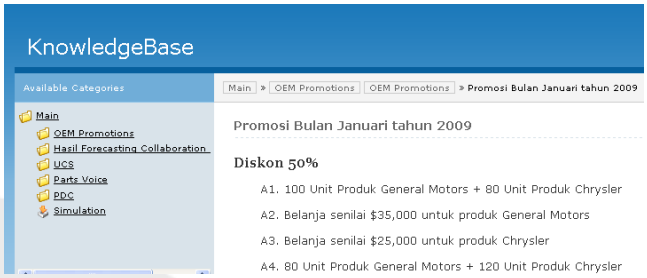


Figure 5. Knowledgebase capture public side

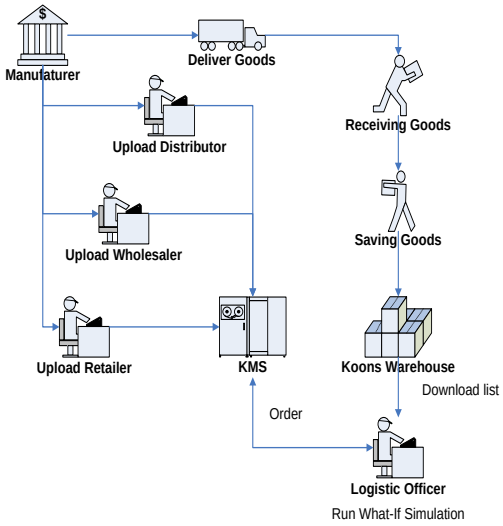


Figure 4. Work flow after implementing KMS

4. IMPLEMENTATION

In order to implement this concept, we use KnowledgeBase trial 30 days from activecompaign.com, the capture of software shown at figure 4.

With this web-based application, we can manage all information and knowledge from manufacturer to retailer, and store all data that needed to achieve the final goal.

In manufacturer side, KMS take the data of yearly promotions program, time based and quantity order based, the straightforward scheme and also the complex scheme.

While in distributor and wholesaler side, they must store the logistic data or conditions of their inventory, and update it regularly. They also can use what-if simulation to get the most high less price to purchase.

In retailer side, they can run what-if simulation and can order electronically using KMS to the wholesaler or distributor.

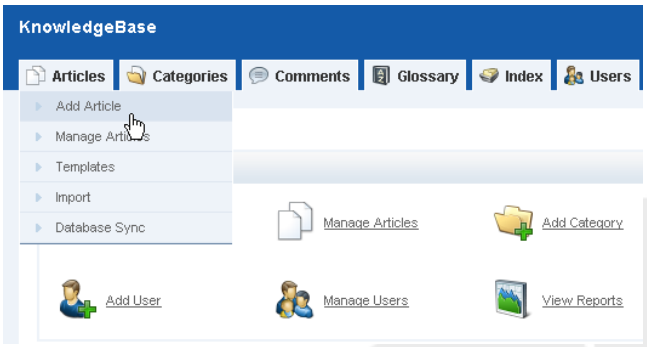


Figure 5. Knowledgebase capture admin side

5. CONCLUSION

In accordance with the original purpose of this research is to build a Knowledge Management System, which is the integration of KM with the SCM, KMS will be used by the Manufacturer, Distributor, Wholesaler and Retailer with KMS.

After doing the research is by KMS will integrate data, information and knowledge from all parts of the Supply Chain, a KMS is shown to help reduce inventory costs, reduce ordering costs, increase sales, and will reduce the sales price into the hands of customers.

We can make conclusions to this work

1. By using the KMS, total sales will increase, as will many who sell Promotions
2. Retailer party also feel disadvantaged by the KMS, as it will reduce costs and reduce order fulfillment costs, because getting a discount from the OEM Promotions
3. End customers also feel disadvantaged because they get reduced prices and will further increase the value of CSI
4. All divisions at internal Supply Chain will synergize each other, because there are no conflicting goals again.

6. REFERENCES

[1] Barnes, Stuart. 2002. Knowledge Management Systems : theory and practice. London : Thomson Learning

[2] Chopra, S., Meindl, P. 2004. Supply Chain Management. New Jersey : Prentice Hall

[3] Franklin, Burt. 1967. Industrial Evolution. USA

[4] Little, S., Ray, T. 2005. Managing Knowledge. London : The Open University

[5] Maier, Ronald. 2007. Knowledge Management System. Austria : Springer

[6] Raghunathan, S., 1999, Interorganizational collaborative forecasting and replenishment systems and supply chain implications, Decision Sciences 4, 1053-1071

[7] Shaw, N. C., Meixell, M. J., Tuggle, F. D., 2002. A Case Study of Integrating Knowledge Management into the Supply Chain Management Process. Proceedings of the 36th Hawaii International Conference on System Sciences. HICSS'03

[8] Seifert, Dirk. 2003. Collaboration Planning, Forecasting and Replenishment. United States of America: Amacom

Indonesian WordNet Sense Disambiguation Using Cosine Similarity and Singular Value Decomposition

Syandra Sari
Faculty of Computer Science
University of Indonesia
Depok INDONESIA
syandra_sari@trisakti.ac.id

Ruli Manurung
Faculty of Computer Science
University of Indonesia
Depok INDONESIA
maruli@cs.ui.ac.id

Mirna Adriani
Faculty of Computer Science
University of Indonesia
Depok INDONESIA
mirna@cs.ui.ac.id

ABSTRACT
This paper describes initial experiments on word sense disambiguation (WSD) for the Indonesian language. WSD is the task of determining the correct sense of a word according to the context it appears in. In these experiments, a number of polysemous words appearing in the Indonesian WordNet are randomly chosen, and Google is used to collect testing context paragraphs. Some well-known vector model metrics are then applied, i.e. cosine similarity and singular value decomposition (SVD) to solve the Indonesian WSD problem. The results are compared against the human judgments of three graduate students who were asked to select the most appropriate definition for each testing context paragraph. The experiment results showed that answers using cosine similarity achieved 62.5% similarity with human answers, whereas SVD achieved 67.5%. Using the Fleiss kappa statistic, cosine similarity-based WSD achieves an agreement of 0.311 with three human judges, whereas SVD is able to achieve 0.350.

Keywords
Word Sense Disambiguation, WordNet, Cosine Similarity, Singular Value Decomposition

1. INTRODUCTION
Word Sense Disambiguation (WSD) is the problem of assigning the appropriate meaning (sense) to a given word in a text or discourse [19]. Automatic WSD is one of the most important open problems in the Natural Language Processing (NLP) field. WSD is used in machine translation, information retrieval, extraction information, text mining and lexicography [3].

There are several methods to solve the WSD problem automatically: methods that use language resources, e.g. heuristic approach [11], semantic similarity measure [1] and Lesk algorithm [14]. Dictionary [13], Wordnet [6], and selectional preference [2] are the example of language resources used for WSD. Another method is the corpus approach, which uses both unsupervised and supervised learning algorithms. Distributional unsupervised is an example of unsupervised method that uses monolingual corpus [16] and translational equivalence is also an unsupervised method using a parallel corpus [5]. There are various supervised approaches, e.g.: bayesian network [7], decision lists [25], *k*-Nearest Neighbor [18] and maximum entropy [20].

Automatic WSD has been developed for many languages throughout the world, e.g. WSD in English [5, 12], Italian [9], Chinese [27], Spanish [10], Japanese [22], Indian [17], Bengali [4], and Korean [26]. There has also been some initial research work into WSD for the Indonesian language using a Naive Bayesian approach [23].

This paper describes our own initial experiments into WSD for the Indonesian language, where well-known vector model metrics such as cosine similarity and singular value decomposition are applied. In particular, the task trying to be solved is determining the correct sense of an Indonesian word from the Indonesian WordNet [28], based on the context the word appears in. The linguistic resources used are Google and our prototype Indonesian WordNet[1].

The paper is organized as follows: Section 2 describes cosine similarity and the singular value decomposition. Section 3 is devoted to careful explanation of the experimental setting. Section 4 reports the set of experiments performed and the analysis of the results obtained. Finally, Section 5 concludes and outlines some directions for future work.

2. METHODS AND RESOURCES

2.1 Cosine Similarity
Cosine Similarity is a similarity metric measured based on vectors. It is usually used for measuring similarities among documents. The set of documents in a collection are viewed as a set of vectors in a vector space, in which there is one axis for each term. It is defined as follows [15]:

$$\mathrm{sim}(d_1, d_2) = \frac{\vec{V}(d_1) \cdot \vec{V}(d_2)}{|\vec{V}(d_1)||\vec{V}(d_2)|}, \quad (1)$$

Where the numerator represents the dot product (also known as the inner product) of vectors $\vec{V}(d_1)$ and $\vec{V}(d_2)$, and the denominator is the product of their Euclidean lengths. The dot product $\vec{x}.\vec{y}$ of two vectors is defined as $\sum_{i=1}^{M} x_i y_i$. Let denote $\vec{V}(d)$ the document vector

[1] http://bahasa.cs.ui. ac.id/iwn

for ***d***, with M components $\overrightarrow{V_1}(d) \ldots . \overrightarrow{V_M}(d)$. The Euclidean length of d is defined as $\sqrt{\sum_{i=1}^{M} \overrightarrow{V_i}^2 (d)}$.

In our work, ***d*** is not a document, but a paragraph. The components of a paragraph are words, and the value of each component can be 0 or 1. If paragraph contains word W_i, the value is 1 and if there is not word W_i in the paragraph the value is 0. The cosine similarity is then applied to show similarities between words appearing in a test paragraph and definitions.

2.2 Singular Value Decomposition

The singular value decomposition (SVD) is a means of decomposing a matrix into a product of three simpler matrices. In this way it is related to other matrix decompositions such as eigen decomposition, principal components analysis (PCA), and non-negative matrix factorization (NNMF). The original and most well known application of SVD in natural language processing has been for latent semantic analysis (LSA). Because of the simple vector representations of terms and documents produced by SVD, SVD also has been widely used for clustering. Outside of strictly linguistic applications, SVD has been used for collaborative filtering, for example in the field of movies. The work reported here is similar to LSA, but we have not yet developed the semantic model on a large external corpus – this is a subject of ongoing work.

Let $\boldsymbol{A} \in R^{nXm}$, n and m are positive integers. The range of A is the subspace of R^n define by $R(\boldsymbol{A}) = \{Ax \mid x \in R^m\}$. The rank of ***A*** is the dimension of $R(\boldsymbol{A})$, and ***A*** is not a zero matrix. The singular value decomposition for ***A*** can be expressed as:

$$\boldsymbol{A} = \boldsymbol{U \Sigma V^T}$$

Where $U \in R^{nXn}$ and $V \in R^{mXm}$ are orthogonal, and $\Sigma \in R^{nXm}$ is a nonsquare diagonal matrix.

$$\Sigma = \begin{bmatrix} \sigma_1 & & & & \\ & \sigma_2 & & & \\ & & \ddots & & \\ & & & \sigma_r & \\ & & & & 0 \\ & & & & & \ddots \end{bmatrix}$$

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0.$$

The entries σ_1, ... σ_r of Σ are uniquely determined, and they are called the singular values of ***A***. The columns of U are orthonormal vectors called right singular vectors of ***A***, and the columns of V are called left singular vectors [24].

In this work, SVD is used for decomposing the matrices built from definition vectors and paragraph vectors.

2.3 Indonesian WordNet

Indonesian WordNet is a database of *synsets* (synonym sets) developed at the Information Retrieval Laboratorium, Faculty of Computer Science, University of Indonesia [28]. WordNet was developed using expand approach so the structure is similar with PWN (Princeton WordNet)[2]. As of now, 1203 unique base concept synsets and 1659 distinct words have been obtained in Indonesian WordNet. There are 2261 semantic relations obtained from synsets including hyponymy and hypernymy. The definitions of 20 randomly chosen polysemous words found in the Indonesian WordNet was used for this experiment.

Table 1. Test words.

No.	Indonesian word	English translation	Number of senses
1	*adat*	tradition	5
2	*biru*	blue	2
3	*bubur*	porridge	3
4	*corong*	funnel	2
5	*erosi*	erosion	2
6	*goni*	gunny	2
7	*hijau*	green	2
8	*jala*	net	2
9	*kecelakaan*	accident	2
10	*kekasaran*	coarseness	2
11	*kokoh*	solid	2
12	*kriminalitas*	criminality	2
13	*lemparan*	throw	2
14	*peningkatan*	enhancement	2
15	*perasa*	sense	2
16	*pertunjukan*	show	2
17	*perusakan*	destruction	2
18	*pesona*	charm	2
19	*teguh*	firm	2
20	*tuan*	master	2

3. EXPERIMENTAL SETUP

Our experiment can be divided into several steps: choosing test words from the Indonesian WordNet for our experiment, searching for testing context paragraphs using Google, and applying Cosine Similarity and Singular Value Decomposition for determining the best definition for a particular test word.

3.1 Choosing polysemous words

In the Indonesian language, there are many words that have more than one meaning or sense. In the current version of the Indonesian WordNet, there are 1947 polysemous words, i.e. words having more than one meaning. From these, 20 words were randomly chosen. Table 1 shows these words.

3.2 Searching Testing Context Paragraphs

WSD is the task of determining the sense of a word appearing in a text. Thus, to apply WSD for Indonesian, test paragraphs containing the above chosen words must be prepared. To achieve this, Google is used by entering each test word as a keyword and

[2] http://wordnet.princeton.edu/

saving the 20 top documents. From these top documents the paragraphs where our target word appears are extracted. Only two paragraphs are selected for the experiment based on the most amount of overlapping words with the definition, or *gloss*, of the

Table 2. Definitions and context for the word *tuan*.

First definition: *keturunan orang mulia-mulia (terutama raja dan kerabatnya); ningrat; orang berbangsa:; sebutan untuk penguasa tertinggi dr suatu kerajaan; sesuatu yg diyakini, dipuja, dan disembah oleh manusia sbg yg Mahakuasa, Mahaperkasa, dsb:; mampu sekali dl bidang ilmu:; orang tempat mengabdi, sbg lawan kata hamba, abdi, budak:; orang yg berpangkat tinggi*
Second definition: *orang yg memberi pekerjaan; majikan; kepala (perusahaan dsb); pemilik atau yg empunya (toko dsb):; orang laki-laki (yg patut dihormati):; persona orang kedua laki-laki ; sebutan kpd orang laki-laki bangsa asing atau sebutan kpd orang laki-laki yg patut dihormati:*
First paragraph: *Kukar **Tuan** Rumah Bupati Kutai Kartanegara (Kukar) Prof Dr Syaukani HR MM dalam Official Handbook menyambut baik kegiatan International Training Workshop on Dispute Settlement Mechanism on Investment itu. Dia menyebut kegiatan itu merupakan suatu kehormatan bagi Pemkab Kukar, yang telah dipercaya selaku tuan rumah. "Menjadi tuan rumah bagi delegasi 14 negara sekaligus adalah pengalaman pertama kami," katanya.*
Second paragraph: *Yakin dong negara kita bisa jadi **tuan** rumah piala duni 2022.Yang penting semua elemen bangsa mau bersatu untuk mewujudkan cita-sita itu.Tul gak, Insya Allah berhasil*

target word senses. The definition of each test word was taken from Indonesian WordNet. Table 2 shows the example of definition and test paragraphs for the word *tuan* (master). These two paragraphs were selected because they have the most number of common words with the two definitions amongst all paragraphs extracted from the top 20 documents.

3.3 Applying Cosine Similarity for WSD

Before applying the cosine similarity metric, the definitions and context paragraphs must first be preprocessed. Firstly, stopwords are eliminated and then stemming is applied to the definitions and paragraphs. The Indonesian stemming was from [20]. A list of words appearing in definitions and paragraphs is then constructed. Tables 3a, 3b and 3c show the list of words for *biru* (blue) from the two definitions and one context paragraph containing the word *biru*. Note that due to the stemming process used, some words can be incorrect, for example: *sepert* should be *seperti*, whereas *elan* should be *telan*.

The lists of all words from both definitions and context paragraphs are combined. Table 3d shows the combined list from the first definition, second definition and a test paragraph of word *biru*. This defines a semantic space for each word, i.e. both Indonesian WordNet senses of biru are distinct vectors in this space. The context paragraph defines a third vector, and the question is to find the nearest sense vector.

Word vectors for each definition based on the list word are created. The vector has two values, 1 means the word is not in the definition and 0 means the word is in definition. Based on Tables 3a and 3b, word vectors such as the following can be constructed:

Biru_1,0:1:1:1:1:0:0:0:0:1:0:0:0:1:1:1:0:0:0:0:0:0:0:1:1:0:1:1:1:0:1:0:1:1:1:1:1:1:1:1:
1:1:1:0:0:1:1:0:1:1:1:0:0:1:0:0:1:0:1:1:0:0:1:0:0:1:1:1:0:1:1:1:0:1:1:0
Biru_2,1:0:0:0:0:1:1:1:0:0:0:1:1:0:0:0:1:0:0:1:1:0:0:0:0:1:0:0:0:0:0:0:0:0:0:0:0:0:0:0:
1:0:0:0:0:0:0:1:0:0:0:1:0:0:0:0:0:0:0:0:0:0:0:0:0:0:0:1:1:0:0:0:0:1:0

A word vector for the testing context paragraph was also created based on the above wordlist. This is an example vector for the context paragraph from Table 3c:

Table 3a. First Definition of word *biru*.

mengandung atau memperlihatkan warna yg serupa warna langit yg terang; lipatan-lipatan pd tepi baju, kain, dsb sbg hiasan; suka berbuat kurang baik (tidak menurut, mengganggu, dsb, terutama bagi anak-anak):; buruk kelakuan (lacur dsb):; nakal; suka usil (mengganggu); pekak atau tuli sementara (krn ditampar, menelan pil kinine, dsb); tidak mau mengindahkan nasihat dsb; keras kepala; nakal; suka mengganggu; kurang ajar; kurang senonoh (kasar) dl bertingkah laku; kurang ajar; nakal:; terasa spt cabai atau merica; tajam atau keras (tt kritik dsb); menyakiti hati (tt perkataan dsb):			
ajar *anak* *baju* *buat* *tingkah* *buruk* *caba* *tampar* *hati* *hias* *kain*	*kasar* *laku* *pala* *keras* *kinine* *kritik* *lacur* *langit* *lipat* *lihat* *elan*	*andung* *ganggu* *indah* *sakit* *rica* *nakal* *nasihat* *pekak* *kata* *pil* *senonoh*	*sepert* *suka* *tajam* *tepi* *terang* *asa* *tuli* *usil* *warna*

Table 3b. Second Definition of word *biru*.

tepung berwarna biru sbg bahan pencampur air pembilas cucian agar warna pakaian menjadi kebiru-biruan, biasanya untuk pakaian berwarna putih; warna dasar yg serupa dng warna langit yg terang (tidak berawan dsb) serta merupakan warna asli (bukan hasil campuran beberapa warna)			
air *asli* *bahan* *awan* *warna*	*biru* *biru* *campur* *cuci* *dasar*	*hasil* *biru* *langit* *pakai* *bilas*	*campur* *putih* *tepung* *terang* *warna*

Table 3c. Test paragraph for word *biru*.

*Tapi tak ada yang membuat mereka tersinggung sekali. Mereka sebetulnya orang yang penyabar, tidak pendendam meskipun hatinya sering luka atau dilukai. Ia melihat dunia ini sebagai satu wilayah yang romantis sekaligus mengandung banyak ranjau yang bisa membahayakan suasana hatinya. Ada kalanya ia tersungkur dan mundur, tapi penyuka warna **biru** biasanya tabah dan mencoba untuk bangkit dan berusaha mencapai apa yang diinginkan.*			
buat *singgung* *sabar* *dendam* *hati* *luka* *luka*	*lihat* *dunia* *satu* *wilayah* *romantis* *andung* *ranjau*	*bahaya* *suasana* *hati* *kala* *sungkur* *mundur* *suka*	*warna* *biru* *tabah* *coba* *bangkit* *usaha* *capa* *ingin*

Test_Par,0:0:0:1:0:0:0:0:1:0:1:0:1:1:0:0:0:1:1:0:0:1:1:0:0:0:1:0:0:1:0:1:0:0:0:0:0:0:
0:0:0:1:1:0:0:0:0:0:0:0:1:0:1:1:0:1:0:0:1:1:1:1:1:1:0:0:0:0:0:0:0:1:0:1:1

Table 3d. List of all words from two definitions and test paragraph for word *biru.*

air	*cuci*	*kritik*	*sakit*
ajar	*dasar*	*lacur*	*satu*
anak	*dendam*	*laku*	*senonoh*
andung	*dunia*	*langit*	*sepert*
asa	*elan*	*lihat*	*singgung*
asli	*ganggu*	*lipat*	*suasana*
awan	*hasil*	*luka*	*suka*
bahan	*hati*	*mundur*	*sungkur*
bahaya	*hati*	*nakal*	*tabah*
baju	*hati*	*nasihat*	*tajam*
bangkit	*hias*	*pakai*	*tampar*
bilas	*indah*	*pala*	*tepi*
biru	*ingin*	*pekak*	*tepung*
buat	*kain*	*pil*	*terang*
buruk	*kala*	*putih*	*tingkah*
caba	*kasar*	*ranjau*	*tuli*
campur	*kata*	*rica*	*usaha*
capa	*keras*	*romantis*	*usil*
coba	*kinine*	*sabar*	*warna*
			wilayah

At this point, the cosine similarity based on these vectors (Eq. 1) can be calculated. For example:

- Cosine Similarity between word vector of Biru_1 and word vector of test paragraph is 0.176
- Cosine Similarity between word vector of Biru_2 and word vector of test paragraph is 0.096

So from these similarity values, it can be said that the word *biru* in test paragraph (Table 3c) agrees with the first definition (Table 3a).

3.4 Applying SVD for WSD

The process for applying SVD is the same as the process used for cosine similarity. The only difference is that after creating word vectors from definitions and test paragraphs, the SVD is applied first, and then used to reduce the rank/dimensionality of the resulting matrix. The process of SVD is as follows:

a. Combine all definition word vectors and a paragraph word vector into one matrix (M)

b. Decompose M using SVD to : U, Σ, and V^T

c. Truncate the dimension of U, Σ, dan V^T . We only use two singular values in matrix Σ. It means we only take two first columns and row from matrix Σ then take two first two columns from matrix U and V^T.

d. Recompose matrix M from new U, Σ, and V^T and get matrix M'

These are seven steps in Matlab for the SVD process:

- [U, Σ, V] = svd(M,0) // decompose matrix M using SVD
- Σ2 = Σ (:,1:2) //take two first columns from matrix Σ
- Σ2B = Σ 2(1:2,:) // take two first rows from matrix Σ2
- U2 = U(:,1:2) // take two first columns from matrix U
- V2 = V(:,1:2) // take two first columns from matrix V
- V2T = V2' // Transpose matrisk V2
- M' = U2 * Σ2B * V2T // recompose

e. Decompose matrix M' to obtain definition vectors and test paragraph vectors.

Here is an example of the definition vectors and test paragraph vectors yielded from the SVD process for the word *biru*:

Biru_1_SVD,0.0784:0.9937:0.9937:0.9834:0.9937:0.0784:0.0784:0.0784:-0.0103:0.9937:-0.0103:0.0784:0.0681:0.9834:0.9937:0.9937:0.0784:-....
Biru_2_SVD,0.0233:0.0784:0.0784:0.2073:0.0784:0.0233:0.0233:0.0233:0.0784:0.1289:0.0233:0.1522:0.2073:0.0784:0.0784:0.0233:0.1289.....
Biru_paragraph_SVD,0.1289:-0.0103:-0.0103:0.9726:-0.0103:0.1289:0.1289:0.1289: 0.9830:-0.0103:0.9830:0.1289:1.1119:0.9726:-0.0103:.....

f. Calculate similarity between each definition vector and test paragraph vector yielded from SVD process using cosine similarity (Eq. 1)

For example:

- Cosine Similarity between word vector of Biru_1_SVD and word vector of test paragraph (Biru_paragraph_SVD) is 0.172
- Cosine Similarity between word vector of Biru_2_SVD and word vector of test paragraph (Biru_paragraph_SVD) is 0.831

So from these similarity values, it can be said that the word *biru* in test paragraph (Table 3c) agrees with the second definition (Table 3b) based on SVD method.

Table 4. Experiment results.

No	Word	Para-graph	Cosine Similarity	SVD	Human assess-ment
1	*adat*	1	Definition 2	Definition 2	Definition 3
		2	Definition 4	Definition 4	Definition 4
2	*biru*	1	Definition 1	Definition 2	Definition 1
		2	Definition 2	Definition 2	Definition 1
3	*bubur*	1	Definition 1	Definition 2	Definition 3
		2	Definition 3	Definition 2	Definition 3
4	*corong*	1	Definition 1	Definition 1	Definition 1
		1	Definition 1	Definition 1	Definition 1
5	*erosi*	2	Definition 2	Definition 2	Definition 2
		2	Definition 2	Definition 2	Definition 2
6	*goni*	1	Definition 2	Definition 2	Definition 2
		2	Definition 1	Definition 2	Definition 2
7	*hijau*	1	Definition 2	Definition 2	Definition 2
		2	Definition 2	Definition 2	Definition 2
8	*jala*	1	Definition 1	Definition 1	Definition 2
		2	Definition 1	Definition 1	Definition 2
9	*kecelakaan*	1	Definition 2	Definition 2	Definition 2
		2	Definition 1	Definition 2	Definition 2
10	*kekasaran*	1	Definition 1	Definition 2	Definition 2
		2	Definition 2	Definition 2	Definition 2
11	*kokoh*	1	Definition 1	Definition 1	Definition 2
		2	Definition 2	Definition 1	Definition 2

12	*kriminalitas*	1	Definition 1	Definition 1	Definition 2
		2	Definition 1	Definition 1	Definition 1
13	*lemparan*	1	Definition 1	Definition 1	Definition 1
		2	Definition 1	Definition 1	Definition 1
14	*peningkatan*	1	Definition 2	Definition 2	Definition 2
		2	Definition 2	Definition 2	Definition 1
15	*perasa*	1	Definition 1	Definition 2	Definition 1
		2	Definition 2	Definition 2	Definition 1
16	*pertunjukan*	1	Definition 2	Definition 2	Definition 2
		2	Definition 2	Definition 1	Definition 1
17	*perusakan*	1	Definition 1	Definition 2	Definition 2
		2	Definition 2	Definition 2	Definition 2
18	*pesona*	1	Definition 2	Definition 2	Definition 2
		2	Definition 1	Definition 1	Definition 1
19	*teguh*	1	Definition 2	Definition 2	Definition 2
		2	Definition 2	Definition 2	Definition 2
20	*tuan*	1	Definition 2	Definition 2	Definition 2
		2	Definition 1	Definition 2	Definition 2

4. RESULTS AND ANALYSIS

According to the explanations in Sections 3.1 and 3.2, only two test paragraphs for each test words were taken for experiment. Thus, 40 test paragraphs are produced from the 20 test words. Table 4 describes the result of our experiment. It shows the target word, the chosen definition using cosine similarity, SVD, and the 'gold standard' human judgment.

This table shows that 28 paragraphs (70%) from 40 test paragraphs give the same result using both methods. Nine words give the same answer from two test paragraphs: *adat, corong, hijau, jala, kriminalitas, lemparan, peningkatan, pesona* and *teguh*. Ten words give only one same answer from two test paragraphs. These are *biru, erosi, goni, kecelakaan, kekasaran, kokoh, perasa, pertunjukan, perusakan,* and *tuan.* The word *bubur* is the only word giving different answer from two test paragraphs.

4.1 Evaluation

To evaluate the automatic WSD, a survey to obtain some human judgments was conducted. Three graduate students were asked to select the most appropriate senses for the 40 test paragraphs. The last column in Table 4 shows the result of this survey. We used this result for evaluating our experiment. The final answer of each paragraph was taken using a majority vote of the chosen definition from two or three out of the judges. Table 4 shows that the cosine similarity method has 25 paragraphs (62.5%) in common with the human answers, whereas SVD has 27 paragraphs (67.5%) in common. There are 21 paragraphs or 52.5% where both methods give the same answers with the human judges. There are nine paragraphs where both methods give different answers from humans. These cases are for the words *adat* (first paragraph), *biru* (second paragraph), *bubur* (first paragraph), *jala* (first and second paragraph), *kokoh* (first paragraph), *kriminalitas* (first paragraph), *peningkatan* (second paragraph), and *perasa* (second paragraph).

The Fleiss kappa statistic [29] for measuring inter-annotator reliability was computed. This value is in the interval [0..1], where 1 shows perfect agreement. We measured three Fleiss kappa values: (i) between the answers of the three human judges, (ii) between the answers of the three human judges and cosine similarity as the fourth judge, and (iii) between the answer of the three human judges and SVD as the fourth judge. For the first, we obtain 0.485, for the second we obtain 0.311, and lastly for the third we obtain 0.350.

Interpreting Fleiss kappa values is very relative to the task that the annotators are required to perform. However, it is generally thought that values of 0.4 and above show moderate to good agreement. From these values, we see that agreement between humans is thus moderate, and that the automatic WSD is not yet able to achieve human performance. However, the Fleiss kappa values also show that the result of SVD is more similar to human judgments than cosine similarity.

5. CONCLUSIONS

This paper reports an investigation into the use of cosine similarity and SVD vector model metrics for WSD in the Indonesian language. Although they are simple methods, our experimental results demonstrated both methods could achieve more than 60% accuracy compared with human judgments for 40 test paragraphs. Of the two, SVD achieved slightly better results than cosine similarity.

In the future we aim to extend the SVD method using a large monolingual corpus or LSA method for WSD problem. Improving the definition of the word and exploring more data about words in Indonesian documents are other ways that may improve the accuracy.

6. REFERENCES

[1] Agirre, E. and Rigau, G. 1996. Word Sense Disambiguation Using Conceptual Density, International Conference On Computational Linguistics. Proceedings of the 16th conference on Computational linguistics - Volume 1, Copenhagen, Denmark, 16 – 22.

[2] Agirre, E. and Martínez, D. 2001. Learning class-to-class selectional preferences. Proceedings of the Conference on Natural Language Learning, Toulouse, France, 15–22.

[3] Agirre, E. and Edmonds, E. 2007. Word Sense Disambiguation Algorithms and Applications, Springer.

[4] Banerjee, S. and Mullick, B.P. 2007. Word Sense Disambiguation and WordNet Technology. Literary and Linguistic Computing, 22: 1-15.

[5] Brown, P.F., Della Pietra, S.A., Della Pietra, V.J., and Mercer, R.L. 1991. Word-sense disambiguation using statistical methods. Proceedings of the 29th Meeting of the Association for Computational Linguistics (ACL), Berkeley, U.S.A., 264–270.

[6] Cañas, A.J., Valerio, A., Lalinde-Pulido, J., Carvalho, M.M., and Arguedas, M. 2003. Using WordNet for Word Sense Disambiguation to Support Concept Map Construction. SPIRE 2003: 350-359.

[7] Escudero, Gerard, Lluís Màrquez and German Rigau. 2000. Naive bayes and exemplar-based approaches to word sense disambiguation revisited, Proceedings of the 14th European Conference on Artificial Intelligence (ECAI), Berlin, Germany, 421–425.

[8] Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. Psychological Bulletin, 76(5):378-382.

[9] Florian, Radu and Richard Wicentowski. 2002a. Unsupervised Italian word sense disambiguation using WordNets and unlabeled corpora. Proceedings of the ACL-02 workshop on Word sense disambiguation: recent successes and future directions, 8: 67-73.

[10] Florian, Radu, Silviu Cucerzan, Charles Schafer, and David arowsky. 2002b. Combining classifiers for word sense disambiguation. Natural Language Engineering 8 (4): 327–341.

[11] Gale, William, Ken Church and David Yarowsky. 1992. One sense per discourse. Proceedings of the DARPA Speech and Natural Language Workshop, New York, U.S.A, 233–237.

[12] Guo, Weiwei and Mona T. Diab. 2009. Improvements to monolingual English word sense disambiguation. Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions, Boulder, Colorado, 64-69.

[13] Krovetz, R and W. B. Croft. 1989. Word sense disambiguation using machine-readable dictionaries, Proceedings of the 12th annual international ACMSIGIR conference on Research and development in information retrieval, Editor: N.J. Belkin dan C.J. van Rijsbergen, June 25-28, 1989, Cambridge, MA.

[14] Lesk, Michael. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. Proceedings of the ACM-SIGDOC Conference, Toronto, Canada, 24–26.

[15] Manning, Christopher D, Prabhakar Raghavan, and Hinrich Schutze. 2008. Introduction to Information Retrieval, Cambridge University Press.

[16] Miller, George and Walter Charles. 1991. Contextual correlates of semantic similarity. Language and Cognitive Processes, 6(1): 1–28.

[17] Mishra, Neetu, Shashi Yadav and Tanveer J. Siddiqui. 2009. An Unsupervised Approach to Hindi Word Sense Disambiguation, Proceedings of the First International Conference on Intelligent Human Computer Interaction, Springer India, 327-335.

[18] Ng, Hwee T. and Hian B. Lee. 1996. Integrating multiple knowledge sources to disambiguate word senses: An exemplar-based approach. Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL), Santa Cruz, U.S.A., 1996, 40–47.

[19] Seo, Hee-Cheol, Hoojung Chung, Hae-Chang Rim, Sung Hyon Myaeng, and Soo-Hong Kim. 2004. Unsupervised word sense disambiguation using WordNet relatives, Computer Speech and Language 18, 253-273.

[20] Suárez, Armando dan Manuel Palomar. 2002. A maximum entropy-based word sense disambiguation system. Proceedings of the 19th International Conference on Computational Linguistics (COLING), Taipei, Taiwan, 2002, 960–966.

[21] Tala, Fadillah . 2003. A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia, M.Sc. Thesis, University of Amsterdam.

[22] Tanaka, Takaaki, Francis Bond, Timothy Baldwin, Sanae Fujita, and Chikara Hashimoto. 2007. Word Sense Disambiguation Incorporating Lexical dan Structural Semantic Information, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), 2007, 477-485.

[23] Uliniansyah, M. Teduh dan Shun Ishizaki. 2006. Word Sense Disambiguation System Using Modified Naive Bayesian Algorithms for Indonesian Language, Information and Media Technologies, Vol. 1, No. 1 pp.257-274.

[24] Watkins, David S. 2002. Fundamentals of Matrix Computations, Second Edition, John Wiley & Sons, Inc., New York.

[25] Yarowsky, David. 2004. Hierarchical decision lists for word sense disambiguation, Computers and the Humanities, 34(2): 179–186.

[26] Yoon, Yeohoon, Choong-Nyoung Seon, Songwook Lee, and Jungyun Seo. 2006. Unsupervised word sense disambiguation for Korean through the acyclic weighted digraph using corpus and dictionary, Information Processing and Management: an International Journal archive, 42:710 – 722.

[27] Zhang, Yuntao, Ling Gong and Yongcheng Wang. 2005. Chinese Word Sense Disambiguation Using HowNet, Lecture Notes in Computer Science Volume 3610, 925-932.

[28] Putra, D.D., Arfan, A., and Manurung, R. 2008. Building an Indonesian WordNet. In Proceedings of the 2nd International MALINDO Workshop. Cyberjaya, Malaysia, 12-13 June 2008.

[29] Fleiss, J. 1971. Measuring nominal scale agreement among many raters. Psychological Bulletin, 76(5):378–382.

Influence of Electronic Media and External Reward Towards Knowledge Sharing Management to Learning Process in Higher Education Institution

Alexander Setiawan
Informatic Engineering Department – Petra Christian University
Siwalankerto 121 – 131 Surabaya 60236 Indonesia
Telp. (031) – 2983455, Fax (031) – 8417658
alexander@peter.petra.ac.id

ABSTRACT
The aim of this research is to examine the factors supporting individuals' knowledge sharing intention. Based on the theory of reasoned action, this study examined influenced of extrinsic reward, and channel richness to knowledge sharing intention. Data was collected using a field study of lecturer and student in higher education institution Yogyakarta. We employ independent sample t-test and PLS (Partial Least Squares) version 2.0. The result show that there isn't perception difference betwen student and lecturer about factors that supporting knowledge sharing intention. The result show that channel richness has played significant part influenced attitude toward knowledge sharing. Extrinsic reward inposed no impact on an individual's attitude toward knowledge sharing. The result from this study confirm the theory of reasoned action. This study also find that subjective norm greater influence knowledge sharing intention than attitude toward knowledge sharing.

Keywords
Theory Of Reasoned Action, Extrinsic Rewards, Channel Richness.

1. INTRODUCTION
In business environment that is full of competition, an organization must have strategy to survive and win the competition in this global environment. Some key factors to success for an organization are determined by its ability to build human resources, taking advantage of information technology and processing knowledge. Human resources here means skills and abilities of an individual in an organization which is equal as how many knowledges exists in that organization (Cabrera & Cabrera, 2005). In order for an organization to have a competitive advantage, individuals in that organization must share their knowledge with other individuals, whether they're in the same organization or not.

A strategy that is based on technology and knowledge is not only needed for business organization, but education organization also need it, as an example is a university. An University is an organization that has a mission to increase the intelligent and life of a nation/race, so that it can become a civilized race, and become a center of knowledge, science, technology, arts, social science, and civilized humanity by conducting a good quality education. Education organization is different from Bussiness organization, a education organization consist of many human resources. Because education organization has many human resources, so the existence of a competition between individuals in that organization is the key to success for a university to increase its human resources quality.

In this research, researcher wants to test a phenomenon know as sharing knowledge, especially for teachers and students in accounting department. The existence of technological improvement in how to process accounting information in USA which pass through four levels, which are *manual system, book keeping machine system, punched card system, and computerized system* influencing a change in how to process accounting information in Indonesia which is based in information technology (Torong, 2000). The existence of this change made the skill requirement of an accountant change. Nowadays, an accountant must have skill in accounting information system, beside manual accounting system. All this time, the accounting education only use manual accounting system. O'Donnell and Moore (2005) in their research also said that many accounting graduate that has no skill in information technology and limited teachers in accounting that understand accounting information system which is based in information technology. Whereas now many money transaction in an organization that is processed using computerized system and based in information technology. By observing phenomenon and to response against change in market needs for a competitive accountant (Amalia, 2006), there exist the need to share knowledge in the field of accounting, especially in a university, which is reputed to educate and produce teachers and students in accounting.

There exist several researchs which test several factors that influence an individual in sharing his/her knowledge, among them are the researchs which is conducted by Bock et al. (2005), Kwok & Gao (2005), Galia (2006) and Burgess (2005).

This writing aim to prove empirically about the influence of external rewards, organizational climate, pressure of social psychology, media diversity, attitude of a person's behavior to share knowledge, and subjective norms of one's intentions to share knowledge at university. It is expected that this paper can also provide empirical validation of growth factors that influence one's intention to share knowledge and are expected to contribute to the

university to allocate resources or to facilitate teachers and students of accounting for intention to increase the sharing of their knowledge. This means that accounting teachers and students are motivated to share knowledge with teachers and with other students.

2. FOUNDATION OF THEORY

2.1 Knowledge Management

To be able to have a competitive advantage, companies are now required to adopt information technology. The development of information technology marked by the emergence of many new innovations. Innovation itself is characterized as a process of change from the three stages of *invention*, *innovation*, and diffusion (King et al., 1994). These innovations are not only influenced by the existence of information technology, but also the incorporation of the process of creation and knowledge transfer. Nonaka (2007) states that the essence of innovation is the creation of knowledge.

There are several advantages possessed by the knowledge that a company able to compete in the global environment full of competition, namely (Stewart, 1997) quoted by Sangkala (2007), which is a non-subtractive, can be owned by many parties, have different funding from other products, rarely have the economic scale, and *unpredictable*.

In order for knowledge can be used and used properly it is necessary to the existence of knowledge management. Some scholars tried to give a definition of knowledge management. Santosu & Surmach (2001) quoted (Sangkala, 2007) tries to provide insight into the management of knowledge as a process in which the company gave birth to the values of intellectual assets and knowledge-based assets. Knowledge management is also defined as a process for obtaining, storing, sharing, and use of knowledge (Davenport & Prusak, 1998 in Bock et al., 2005). From the above sense can be concluded that knowledge management is an approach to managing intangible assets in this case the intended knowledge so that the organization can have a competitive advantage compared to other organizations.

2.2 Theory of Reasoned Action (TRA)

This Theory of Reasoned Action (TRA) was developed by Icek Ajzen and Martin Fishbein. This theory explains how a person's behavior is influenced by one's intentions to do something. In accordance with its name as the theory of reasoned action, this theory reveals that basically a person behaves in a way that is consciously and based on a specific considerations.

Both the considerations of the gained outcome and taking into account the available information (HARTONO 2007). Generally, the theory of reasoned action can be described as figure 1. From the pictures figure 1 can be explained that a person's behavior (actual behavior) is influenced by one's intentions toward the behavior (behavioral intention). According to Hartono (2007), behavioral intentions and behavior are two different things. Behavior intention or intention (behavioral intention) is the desire to do the behavior, so in this case intention is still not behavior. A person's intention towards a behavior is influenced by two main determinants, which are attitudes toward behavior and subjective norm. Attitude is determined by a strong conviction about the behavior. While the subjective norm is determined by a belief that individuals or particular groups approve or not to a specific action (Hartono, 2007).

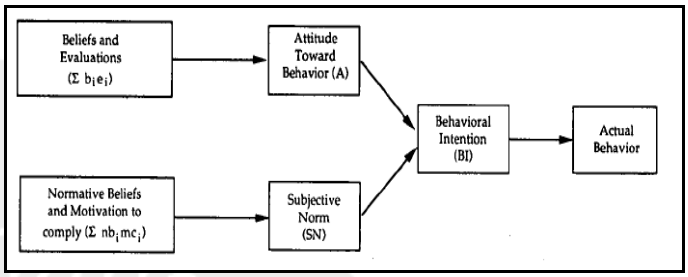


Figure 1. Theory of Reasoned Action

3. DISCUSSION AND ANALYSIS

3.1 Internal Reward Hypothesis

Social exchange theory which suggests that an individual have a desire to interact with other individuals because based on the individual's personal desires which usually cost analysis of the interaction benefit (Blau (1964) quoted Cabrera Cabrera. 2005). Based on the theory, it can be concluded that a person's behavior will be influenced by perceptions of the benefits to be gained from such behavior. In order to share knowledge, provision of benefits is expected that a person can be motivated to share knowledge.

There are several research that investigated the effects of external rewards to a person desire to share knowledge. Galia (2006), Moon & Park (2002) and Burgess (2005) examine the influence of factors external motivation in a person's behavior in the sharing of knowledge. The test results showed that external motivation positively influence employee behavior in the sharing of knowledge.

As in previous research, Bock et al. (2005) and Kwok & Gao (2006) also examined the relationship between external rewards for someone behavior in the sharing of knowledge. Both this study uses the theory of reasoned action (TRA) as the basis to test someone intention in sharing knowledge. The result of research revealed that the awards as a form of external motivation and the negative effects are not significant for someone attitude in sharing knowledge.

Based on social exchange theory and the results of previous research, the first hypothesis formulated buffers as follows:

H1 : External rewards associated with attitudes toward the behavior of someone to share knowledge.

3.2 Media Diversity Hypothesis

There are several research that examine the influence of media diversity in the willingness to share knowledge. Research conducted by Muray and Peyrefitte (2007) and Kwok and Gao (2006) examined a variety of media communications, meetings, and training in order to motivate knowledge sharing. His research shows that there is a positive relationship between the diversity of the media to attitudes toward the behavior of someone to share knowledge.

Based on the results of previously conducted research, the hypothesis can be formulated as follows:

H2 : Media diversity to share knowledge related to positive attitudes toward the behavior of someone to share knowledge.

3.3 Theory of Reasoned Action Hypothesis

This theory explains how someone act is influenced by his/her intention to do something. There is one research that examined the relationship between the influence of attitudes toward someone behaviour to share knowledge with the intention of someone to share knowledge, the influenced of subjective norm to share knowledge with someone intention to share knowledge is researched by Bock et al. (2005). The result of the research is showed that there is positive relation between someone attitude to share knowledge and subjective norm with someone intention to share knowledge. In addition to Bock et al. (2005) also examined the influenced of subjective norm toward attitude to share knowledge. This based on argument assumption from Lee (1990) which quoted by Bock et al. (2005) is suggest that an individual can be motivated to be positive in knowledge sharing when that situation is fit with group norm. On this research Bock at al (2005) discovered there is a positive relation between subjective norm of knowledge sharing with attitude toward behaviour of knowledge sharing.

Based on the theory of reasoned action (TRA) and the previous reasrch that conducted, the hypothesis can be formulated as follows:

H3 : Subjective norm to knowledge sharing is positively associated with the attitude toward behaviour to knowledge sharing.

H4 : Attitude toward behaviour to knowledge sharing is positively associated with someone intention to knowledge sharing.

H5 : Subjective norm to knowledge sharing is positively associated with someone intention to knowledge sharing.

This type of research is hypothesis testing research. Hypothesis figure 2. which want to tested in this research is the influenced of external rewards, organitation of climate, social phsycology tension, media diversity, an attitude toward someone behaviour to knowledge sharing, and the subjective norms toward someone intention to knowledge sharing. Methods of data collection in this research is the survey by using the technique of distributing questionnaires to the respondents figure 3.

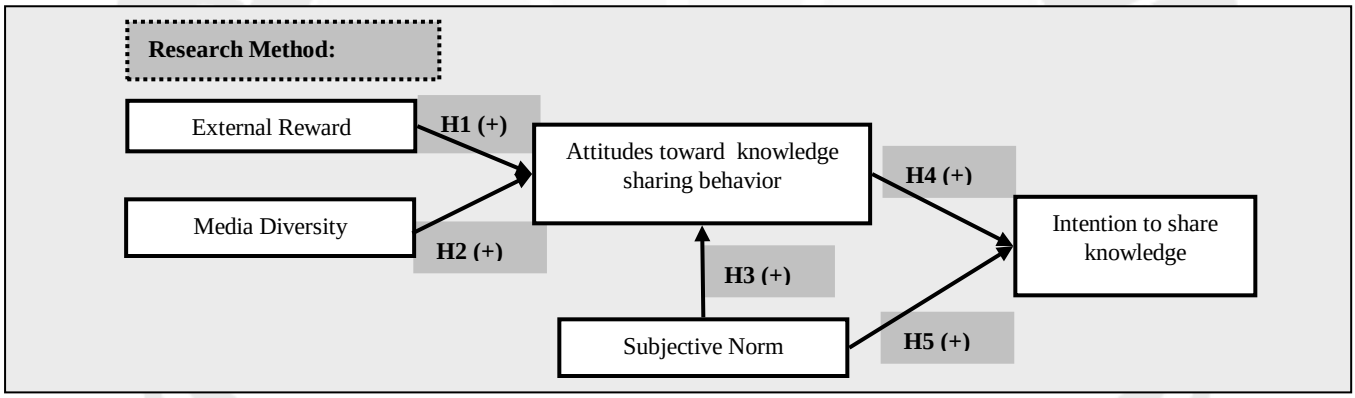


Figure 2. Research Method (1)

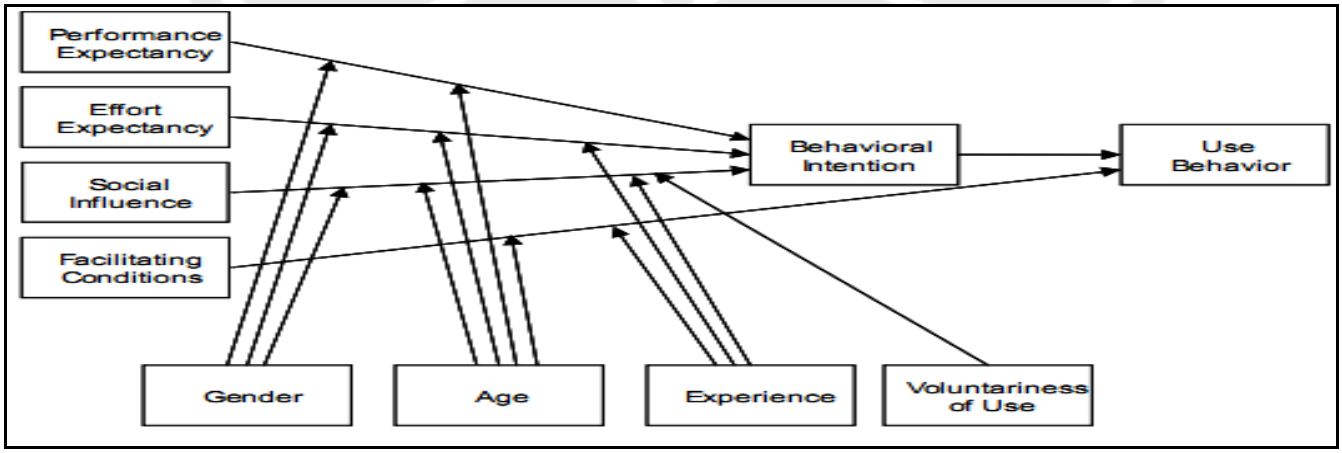


Figure 3. Research Method (2)

3.4 Validity and Reliability Test

Validity test is related to the precision measuring instrument to do its work to reach target (Jogiyanto, 2004). Validity test is divided into two groups namely the content validity and construct validity. The validity of measuring the extent to which the content items in the instrument that measured characteristics represent the attributes to be measured. To ensure content validity, researchers conducted a review of research questionnaires to a friend as well as research respondents during the preliminary tests carried out. Construct validity indicates how well the results obtained from the use of a measure in accordance with the theories used to define a construct (Jogiyanto, 2004). Construct validity was assessed through convergent validity and discriminant validity. Convergent validity is judged by the correlation between score items/indicators with it's construct score. individual indicators considered valid if the correlation value above 0.7 (Ghozali, 2006). Following table 1 convergent validity test results from the data obtained.

Table 1. Validity Test

Variabel	*Factor Loading*
Reward	0.958
Channel	0.869
Attitude	0.910
Norm	0.862
Intention	0.799

Source : Data Processed

As measured by using a construction validity convergent validity test has also been measured discriminant validity. Discriminant validity can be measured by comparing the crossloading between indicator with it's construct (Ghozali, 2006). The following table 2 and the correlation between the construction of indicators.

Table 2. Validity Discriminant Test

	attitude	channel	Intent	Norm	Reward
Attitude1	0.928759	0.431286	0.329921	0.392481	-0.044929
Channel	0.353880	0.863455	0.174409	0.312140	-0.112330
Intention	0.222601	0.196849	0.779668	0.513577	-0.062163
Norm	0.287726	0.317093	0.431331	0.864368	0.026695
Reward	-0.05147	-0.08824	-0.07026	-0.01864	0.995248

Source : Data Processed

Reliability is the level of how much a gauge to measure the stable and consistent (Jogiyanto, 2004). Research instrument is said to have high reliability value if the results of the implementation of various measures on the same subject obtained relatively similar results, for aspects that are measured in the subject have not changed.

Reliability of measurement can be done by looking at the value of composite reliability (Ghozali, 2006) and cronbach's alpha (Nunnally, 1978 in Jogiyanto (2004). A construct is considered reliable if it's reliability composite score above 0.7 (Chin, 2006 cited Bock et al (2005) and values cronbach's alpha above 0.7, but the scale of development research is acceptable loading 0,5-0,6 (Ghozali, 2006). The following Cronbach's alpha values and the composite reliability of each building.

Table 3. Reliability Test

Construct	*Composite Reliability*	*Cronbach's Alpha*
Reward	0.842715	0.812431
Channel	0.873459	0.759954
Attitude	0.918912	0.916988
Norm	0.854604	0.759529
Intention	0.831671	0.775615

Source : Data Processed

3.5 Testing Research

There are two types of tests in this research is to use a test average of different tests and test research models. Average difference test in this research using SPSS 12 (*Statistical Program for Social Science*). While to test the relationships between research variables used PLS 2.0 (*Partial Least Square*).

In this research used PLS analysis methods because the research model used in this research complex. PLS analysis methods are also deemed to have included multiple regression analysis, path analysis, and canonical correlation (Chin, 2000).

In this research, testing the average difference is used to examine differences in faculty and student perceptions of the factors that influence one's intention to share knowledge. The following test results using different test average at table 4.

Table 4. Independent Sample Test for faculty and student

Variable	t-test	significance
Reward	1,921	0,049
Channel	1,595	0,093

Source : Primary Data Processed

From the table above can be seen that there is no real difference between faculty and student perceptions related to several factors that affect a person's attitude in sharing knowledge, including external rewards (REWARD) and diversity of the media (CHANNEL).

Hypothesis testing in this study using PLS (Partial Least Squares). PLS is used in hypothesis testing in this study using the 2.0 version of PLS. The following figure 4 research hypothesis testing results.

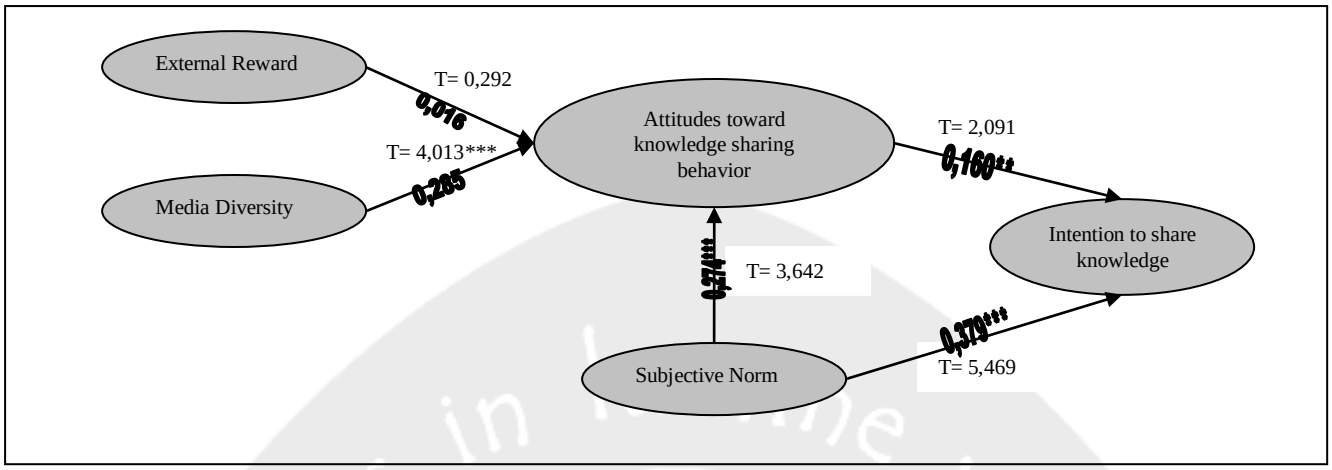


Figure 4. Hypothesis Testing Result

4. CONCLUSION

The purpose of this research is to prove empirically the factors that influence someone to share knowledge that consisting of external rewards, media diversity, sharing knowledge and attitudes toward the behavior of knowledge sharing and subjective norms toward someone intention to share knowledge. Test the average differences that have been performed to determine the different perceptions of students and faculty indicate that there is no significant difference in perceptions regarding the factors that influence one's intention to share knowledge between lecturers and students. Data processing results concluded that external rewards do not significantly affect someone attitude in sharing knowledge. The results of data analysis concludes that media diversity is the main factor affecting the attitude of sharing knowledge with faculty and students. The test results also concluded that the purpose of sharing someone knowledge has been influenced by subjective norms than by attitudes toward knowledge sharing behavior. This is due to the culture of which the place of this research is conducted has a culture of collectivism, so the behavior, largely determined by the rules and the wishes of the community in general than the personal desire for an individual.

5. REFERENCES

[1] Bowman, M., Debray, S. K., and Peterson, L. L. 1993. Reasoning about naming systems. ACM Trans. Program. Lang. Syst. 15, 5 (Nov. 1993), 795-825. DOI= http://doi.acm.org/10.1145/161468.161471.

[2] Ding, W. and Marchionini, G. 1997 A Study on Video Browsing Strategies. Technical Report. University of Maryland at College Park.

[3] Fröhlich, B. and Plate, J. 2000. The cubic mouse: a new device for three-dimensional input. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (The Hague, The Netherlands, April 01 - 06, 2000). CHI '00. ACM Press, New York, NY, 526-531. DOI= http://doi.acm.org/10.1145/332040.332491

[4] Tavel, P. 2007 Modeling and Simulation Design. AK Peters Ltd.

[5] Sannella, M. J. 1994 Constraint Satisfaction and Debugging for Interactive User Interfaces. Doctoral Thesis. UMI Order Number: UMI Order No. GAX95-09398., University of Washington.

[6] Forman, G. 2003. An extensive empirical study of feature selection metrics for text classification. J. Mach. Learn. Res. 3 (Mar. 2003), 1289-1305.

[7] Brown, L. D., Hua, H., and Gao, C. 2003. A widget framework for augmented interaction in SCAPE. In Proceedings of the 16th Annual ACM Symposium on User interface Software and Technology (Vancouver, Canada, November 02 - 05, 2003). UIST '03. ACM Press, New York, NY, 1-10. DOI= http://doi.acm.org/10.1145/964696.964697

[8] Y.T. Yu, M.F. Lau, "A comparison of MC/DC, MUMCUT and several other coverage criteria for logical decisions", Journal of Systems and Software, 2005, in press.

[9] Spector, A. Z. 1989. Achieving application requirements. In Distributed Systems, S. Mullender, Ed. Acm Press Frontier Series. ACM Press, New York, NY, 19-33. DOI= http://doi.acm.org/10.1145/90417.90738

Information and Technology Outsourcing Vendor Selection: An Integrative Literature Review

Jimmy
Universitas Surabaya
Raya Kalirungkut, Surabaya
Indonesia
+62 31 298 1395
jimmy@ubaya.ac.id

ABSTRACT

Client-vendor relationship has been considered as a critical and even most critical among other ITO key success factors. For a successful relationship, the client should firstly clarify its expectations as a foundation to assess and select the best vendor proposition to fulfill the expectations. For that reason, this paper is focused to propose criterions to select the best available IT outsourcing vendor to deliver the expectations. Started with grouping the expected ITO potential benefits which can be pursued by the ITO clients, this paper will then propose a list of criterions to evaluate the best vendor's propositions and capabilities to fulfill the benefits. With a clear understanding on each of the criterions, it is hoped that a client organization could have better and richer information to properly consider which vendor to be engaged.

Keywords

IT Outsourcing, Vendor, Selection

1. INTRODUCTION

Information and Technology Outsourcing (ITO) can be defined as a contractual arrangement to source all or part of the organization's IT and/or IS functions from one or more external service providers (Goles & Chin, 2005, p. 49). For various motivations, outsourcing has become a common practice to procure the provision of some or all IT requirements and, in early 2000, has been adopted by most companies in Australia, USA and UK (Cullen & Willcocks, 2003, p. xviii). Although each ITO contract often involve a huge amount of money (e.g. on April 2008, EDS won three ITO contracts worth from US$ 74 million to US$391 million (EDS, 2008)), current researches estimated that about 25 - 30% of the relationships is or will be a failure (Lacity et al, 2008; Goles & Chin, 2005). Such disconcerting facts (i.e. high cost with high risks) has emerged the need of a better knowledge on how to successfully deliver the ITO expectations.

Among diverse ITO key success factors proposed in current literatures, client-vendor relationship has been regarded by most researchers as a critical, if not the most critical, factor to be managed for ITO to success (Dibbern et al., 2004; Fisher et al., 2008; Goles & Chin, 2005; Goo & Nam, 2007; Gottschalk & Solli-Sæther, 2005, Lacity et al., 2008).

Prior to successfully manage an ITO relationship, it is sensible that the client should firstly able to find the right partner(s) to deliver their outsourcing goals. Moreover, study by Cullen & Willcocks (2003) confirms that majority of problems encountered during the contract term were caused by the supplier. Thus, selecting the right vendor must also be regarded as a key success factor (Dibbern et al., 2004; Feeny et al., 2005; Fisher et al., 2008; Gonzalez et al., 2005; Willcocks & Lacity, 2006) which should be appropriately conducted before engaging the supplier(s) under contractual agreement (Cullen & Willcocks, 2003).

Acknowledging the knowledge requirement, this paper will focus on proposing a list of IT outsourcing supplier selection criteria. In order to do so, as suggested by various literatures (Berry, 2006; Dominguez, 2006; Feeny et al., 2005), this paper will firstly synthesize client's ITO expected outcomes as a foundation for later decisions in assessing and selecting outsourcing vendor(s). Further, this paper will also discuss variety of decisions which should be made prior to vendor selection process.

The remainder of this paper will be structured into five main sections. Started with findings from current literatures regarding client's outsourcing expectations, this paper will then synthesize findings on various options regarding vendor selection, which includes vendor configuration and vendor evaluation criteria. The following section will then Discussion on the findings will then be discussed on the fourth section. Finally, the last section will present the conclusion and the implications of this paper.

2. IT OUTSOURCING EXPECTATIONS

Before engaging any outsourcing relationship, it is imperative that client's outsourcing expectations from the firm level view should be clearly stated. Such decision is crucial since it will be the foundation to produces decisions for the rest of the outsourcing lifecycle including deciding which potential vendor is the best vendor to deliver the expected outcomes (Cullen et al., 2007b). However, consensus on company's expectation is exceptionally difficult, if not impossible, to achieve (Hirschheim & Lacity, 2000; Lacity & Rottman, 2008). Complexity to determine the ITO expectations aroused when the various stakeholders involved in the deal started to bring their own agenda which often conflict other stakeholders' agenda (Dibbern et al., 2004; Hirschheim & Lacity, 2000; Lacity & Rottman, 2008). Further, as argued by Cullen et al. (2007b), client's expectation tend to change overtime, thus client's expectations should be carefully managed by both parties (client and supplier) and changes in the expectations should accordingly alter the management and measurement of the outsourcing practice.

Regarding ITO expected outcomes, numerous literatures have been written to discuss various ITO benefits pursued by the client as their underlying motivation to outsource IT. Different authors tend to give different names to explain the similar ITO benefits and some authors explode single benefit into several similar outcome

terms. For example, Cullen et al. (2007b) suggest "remedy for poor performance" and "improve service" outcomes which can be merged into one benefit: "improve service". By synthesizing findings from current literatures, intended outcomes pursued by client organizations can be clustered into thirteen IT outsourcing benefits. Table 1 lists the ITO benefits in no particular order along with the supporting literatures.

Table 1. ITO potential benefits for the client organization

Expected ITO outcome	Supporting literature
Acquire best value for money	Ang & Straub, 1998; Brown & Wilson, 2005; Cullen et al., 2007b; Dibbern et al., 2004; Dominguez, 2006; Lacity & Rottman, 2008; Lacity & Willcocks, 1998; Schniederjans et al., 2005; Seddon et al., 2002; Sparrow, 2003
Improve financial control	Cullen et al., 2007b; Dibbern et al., 2004; Dominguez, 2006; Lacity & Rottman, 2008; Lacity & Willcocks, 1998; Schniederjans et al., 2005; Seddon et al., 2002; Sparrow, 2003
Acquire cash	Brown & Wilson, 2005; Cullen et al., 2007b; Dominguez, 2006; Schniederjans et al., 2005
Improve service	Brown & Wilson, 2005; Cullen et al., 2007b; Dibbern et al., 2004; Dominguez, 2006; Lacity & Rottman, 2008; Lacity & Willcocks, 1998; Schniederjans et al., 2005; Seddon et al., 2002; Sparrow, 2003
Obtain service not available internally	Cullen et al., 2007b; Dibbern et al., 2004; Dominguez, 2006; Seddon et al., 2002; Lacity & Rottman, 2008; Sparrow, 2003; Brown & Wilson, 2005; Schniederjans et al., 2005
Downsizing	Cullen et al., 2007b; Dibbern et al., 2004; Lacity & Willcocks, 1998
Improve work practice flexibility	Cullen et al., 2007b; Dibbern et al., 2004; Lacity & Rottman, 2008; Lacity & Willcocks, 1998; Schniederjans et al., 2005; Seddon et al., 2002; Sparrow, 2003
Concentrate on core competencies	Brown & Wilson, 2005; Cullen et al., 2007b; Dibbern et al., 2004; Dominguez, 2006; Lacity & Willcocks, 1998; Schniederjans et al., 2005; Seddon et al., 2002; Sparrow, 2003; Quinn & Hilmer, 1994
Focus internal IT on high value activities	Brown & Wilson, 2005; Cullen et al., 2007b; Dibbern et al., 2004; Dominguez, 2006; Lacity & Rottman, 2008; Lacity & Willcocks, 1998; Schniederjans et al., 2005; Sparrow, 2003
Support organizational change	Brown & Wilson, 2005; Dibbern et al., 2004; Dominguez, 2006; Lacity & Rottman, 2008; Lacity & Willcocks, 1998; Schniederjans et al., 2005; Sparrow, 2003
Satisfy external or internal mandate	Dibbern et al., 2004; Dominguez, 2006; Lacity & Willcocks, 1998
Evaluate in-house IT function	Lacity & Willcocks, 1998
Minimize risk	Brown & Wilson, 2005; Schniederjans et al., 2005; Willcocks & Lacity, 2006

Among the many potential benefits offered by the ITO practice, client organization could choose to focus on pursuing one or more benefits as their goals. However, ITO managers should decide it carefully since achievement in one outcome could incorporate both negative and positive impacts on service delivery performance, e.g. cost reduction in IT could lead to reduction in quality of service (Hirschheim & Lacity, 2000, p. 107). Furthermore, research by Cullen et al (2007b) shows that as clients gain more experience in managing outsourcing, they tend to focus on fewer benefits than before. Learning from the experienced, it is suggested not to focus on achieving many benefits at one time.

3. VENDOR CONFIGURATION

Once the outsourcing expectations have been decided, client could use the decision to start defining their ITO configuration. ITO configuration as defined by Cullen et al. (2005) is "a high-level description of the set of choices the organization makes in crafting its IT outsourcing portfolio". Configuring ITO is a crucial process and has been regarded as a key factor to determine the outsourcing success (Fisher et al., 2008; Gonzalez et al., 2005; Seddon & Cullen, 2007; Willcocks & Lacity, 2006). Moreover, set of decisions made in the configuration will determine types of management and decisions required for the rest of outsourcing lifecycle (Cullen et al., 2007a).

Although, many decisions regarding ITO configuration are important and must be considered in determining vendor selection criteria, this paper will focus on discussing vendor configuration which is specifically address the supplier not the client relationship with suppliers. Decision on "supplier grouping" and "supplier location" are regarded as vendor configuration's attributes since they determine how many and where should client look for their prospective suppliers.

3.1 Supplier Grouping

Borrowing Cullen et al's (2005) term, supplier grouping defines the number of supplier involved in an outsourcing transaction. In addition, decision on supplier grouping also includes type of arrangement between client and their suppliers.

In term of number, taxonomy of outsourcing relationship proposed by Gallivan & Oh (1999) can be used to comprehensively represent all possible cardinality between client and supplier. The authors claimed that fundamentally, there are four distinct client-vendor cardinality types: simple dyadic (one client – one vendor), multi vendor (one client – many vendors), co-sourcing (many clients – one vendor) and complex (many clients – many vendors).

In term of arranging multiple vendors, Cullen et al. (2005) promote three possible methods which often used in practice: "prime

contractor", "best-of-breed" and "panel". Prime contractor is used to enable access to several providers' best services while streamlining the outsourcing process by selecting one head supplier to manage the other subcontractors. Best-of breed is used by directly managing several providers to promote vendor competition and flexibility in switching vendor. Lastly, panel is used to create ongoing competition among several preferred suppliers to win client's various orders.

There is no fast and easy rule to determine how many suppliers should be used and how to manage them. Each available option has their own rationale and has been used in many successful and failure ITO practices. However, it is recommended that client should not source all of their IT functions only from one provider. Research by Lacity & Willcocks (1998) shows that selective sourcing has a higher possibility to success than total insourcing (source more than 80% of IT budget internally) or total outsourcing (outsource more than 80% of IT budget to one supplier). Moreover, increasing the amount of work to a particular vendor will eventually increase the potential vendor switching cost which could have considerable impacts when the relationship ends (Kaiser & Hawk, 2004).

3.2 Supplier Location

As telecommunication technology become faster and cheaper, nowadays, options to source IT services from suppliers located in anywhere around the globe have become feasible. Defined as the practice to outsource IT services to service provider located in different continent (Rottman & Lacity, 2008, p. 259), offshore outsourcing has become an alternative for company to source their IT service at lower cost than outsourcing IT to domestic vendors or their own in-house IT.

Other than reducing cost, offshore outsourcing has been claimed to potentially able to deliver other strategic benefits such as quality, speed and agility (Willcocks & Lacity, 2006). However, managers should also be aware of additional risks and challenges entailing the offshoring practice. Research shows that more than 50% offshore outsourcing practices failed to reduce IT cost (Dominguez, 2006, p. 23). Rottman & Lacity (2008) note that the following factors have caused more challenges in offshore ITO practice: time zone differences, the need for more controls, cultural differences, different requirement definition used, and difficulties in managing dispersed teams.

4. VENDOR EVALUTION CRITERIA

Once informed with both expected outcomes and available options in vendor configuration, client could start designing vendor evaluation criteria which specifically fits their needs. It has been claimed that the key to select the right vendor is to find vendors who possess the right capabilities to deliver client's ITO expectations (Feeny et al., 2005; Hyder et al., 2006). Although it is mandatory to choose suppliers with the right capabilities, their capabilities should not be the sole consideration made in selecting vendors. Since, in some cases, not all vendors are willing to deliver their best capabilities to the client (Willcocks et al., 2007). Therefore, client should evaluate both vendor propositions (the bid) and vendor capabilities (the bidder) to determine which vendor proposition are the best deal for them (Cullen & Willcocks, 2003).

4.1 Vendor Proposition – THE BID

Evaluating proposition is used to evaluate details of vendors' proposal which includes, for example, price, payment mechanism, and any value along with its inherent risks. Table 2 lists general criterions to evaluate vendor's proposal with no particular order.

Table 2. Criterions to evaluate vendor proposition

Criteria Component	Supporting literature
Solution	Sparrow, 2003; Cullen & Willcocks, 2003; Brown & Wilson, 2005; Berry, 2006
Vendor objectives	Sparrow, 2003; Willcocks et al., 2007
Cost and best value for money	Sparrow, 2003; Cullen & Willcocks, 2003; Dominguez, 2006; Brown & Wilson, 2005; Berry, 2006
Staffing approach	Sparrow, 2003; Cullen & Willcocks, 2003
Risk and risk management approach	Sparrow, 2003; Cullen & Willcocks, 2003; Schniederjans et al., 2005; Berry, 2006
Transition approach	Sparrow, 2003; Cullen & Willcocks, 2003
Account management approach	Cullen & Willcocks, 2003; Dominguez, 2006; Fink & Shoeib, 2003
Financial approach	Sparrow, 2003; Cullen & Willcocks, 2003

4.2 Vendor Capabilities – THE BIDDER

Vendor capabilities concerns with vendor capability to fulfill their proposal and deliver the client's expectations. In addition to vendor capabilities evaluation, client should also consider vendor's experience and reputation in each capability (Brown & Wilson, 2005; Cullen & Willcocks, 2003; Dominguez, 2006; Fink & Shoeib, 2003). Table 3 list criterions which can be used to evaluate vendor capabilities.

Table 3. Criterions to vendor's capabilities

Vendor Capability	Supporting literature
Relationship management	Cullen & Willcocks, 2003; Sparrow, 2003; Hyder et al., 2006; Feeny et al., 2005; Dominguez, 2006; Brown & Wilson, 2005; Fink & Shoeib, 2003
Domain expertise	Cullen & Willcocks, 2003; Sparrow, 2003; Hyder et al., 2006; Feeny et al., 2005; Brown & Wilson, 2005; Berry, 2006; Fink & Shoeib, 2003
Leadership	Cullen & Willcocks, 2003; Feeny et al., 2005; Berry, 2006
Contract management	Cullen & Willcocks, 2003; Hyder et al., 2006; Feeny et al., 2005; Brown & Wilson, 2005; Berry, 2006
Financial	Cullen & Willcocks, 2003; Sparrow,

performance	2003; Feeny et al., 2005; Brown & Wilson, 2005; Berry, 2006; Fink & Shoeib, 2003
Staff management	Cullen & Willcocks, 2003; Sparrow, 2003; Hyder et al., 2006; Feeny et al., 2005; Dominguez, 2006; Brown & Wilson, 2005; Berry, 2006
3rd party management	Cullen & Willcocks, 2003; Sparrow, 2003; Dominguez, 2006; Brown & Wilson, 2005
Technical proficiency	Cullen & Willcocks, 2003; Sparrow, 2003; Hyder et al., 2006; Feeny et al., 2005; Brown & Wilson, 2005; Fink & Shoeib, 2003
Process re-engineering	Cullen & Willcocks, 2003; Sparrow, 2003; Feeny et al., 2005; Brown & Wilson, 2005; Berry, 2006
Knowledge management	Hyder et al., 2006; Oshri et al., 2007
Risk management	Sparrow, 2003; Hyder et al., 2006; Brown & Wilson, 2005; Berry, 2006
Performance improvement	Cullen & Willcocks, 2003; Hyder et al., 2006; Levina & Ross, 2003; Feeny et al., 2005; Dominguez, 2006; Brown & Wilson, 2005
Service delivery	Hyder et al., 2006; Feeny et al., 2005; Dominguez, 2006; Berry, 2006; Fink & Shoeib, 2003
Service transfer	Hyder et al., 2006; Feeny et al., 2005; Dominguez, 2006; Berry, 2006

5. DISCUSSION

Based on the expectations, client should determine which vendor capabilites are important and which are not important. While some capabilities might not be applicable at all, some others could be very crucial. For example, vendor's capability and approach to manage staff are only applicable when the client transfers their staffs as part of the outsourcing arrangement. On the other hand, when the expectation is to reduce cost then it is mandatory that the cost proposed by the supplier is lower than the internal cost. Therefore, weighting factor should be assigned for each criterion before it is used to evaluate the potential supplier.

Based on its necessity, each criterion could then be clustered into two distinct groups: mandatory and non-mandatory criteria. Mandatory criteria consist of crucial criterions that must be met by the supplier. Since these criterions are not negotiable, results from evaluating the supplier using mandatory criteria should be in binary answer only (e.g. yes/no or true/false). Mandatory criterions can be used to decide supplier's eligibility for further evaluation.

Once the supplier proposition pass all mandatory evaluation, client could start evaluate the potential supplier against the non-mandatory criteria. Non-mandatory criteria consist of negotiable criterions where each criterion holds a particular weighting factor based on its necessity to help determine its relative value to the arrangement (Cullen & Willcocks, 2003; Willcocks et al., 2007).

Although the weighting process could eventually determine which supplier(s) is the right supplier, the actual challenge lies on determining which criteria is important and if it is, how much weight is appropriate to describe its relative value for the outsourcing arrangement success (Cullen & Willcocks, 2003). Further, criterions listed above are a very high level definition which should be further detailed using client's specific situation. For example, client could break down relationship management into several criteria such as ability to speak a particular language or 24 hours support.

6. CONCLUSION

Client-vendor relationship has been considered as a critical and even most critical among other ITO key success factors. To achieve such successful relationship, the client should firstly clarify its expectations and use it as a foundation to assess and select the best vendor proposition to fulfill the expectations.

By integrating findings from various literatures, ITO client expectation can be clustered into thirteen groups: acquire best value for money, improve financial control, acquire cash, improve service, obtain service not available internally, downsizing, improve work practice flexibility, concentrate on core competencies, focus internal IT on high value activities, support organizational change, satisfy external or internal mandate, evaluate in-house IT function, and minimize risk. While there are plenty of significant benefits that can be acquired from an ITO arrangement, it is advised that ITO client should focus on only a few which are considered as the most critical advantages to pursue.

Based on the expectations, ITO client could start to consider the number of supplier(s) to be engaged and where should they look for their prospective suppliers. With clear understanding on both ITO expected outcomes and available vendor options, next step in sequence is to design a list of vendor evaluation criteria which ideal to deliver the expectations. Considering the possibility of an ITO vendor not willing to deliver the full capabilities, client organization should evaluate both vendor propositions (the bid) and vendor capabilities (the bidder) to determine which vendor proposition are the best deal for them (Cullen & Willcocks, 2003). Literatures show that vendor proposition can be evaluated using based on the offered solution, vendor objectives, cost and best value for money, staffing approach, risk and risk management approach, transition approach, account management approach, and financial approach.

As for the vendor's capabilities, fourteen criterions exist to be carefully considered: relationship management, domain expertise, leadership, contract management, financial performance, staff management, 3rd party management, technical proficiency, process re-engineering, knowledge management, risk management, performance improvement, service delivery, and service transfer. Based on its necessity, each criterion could then be clustered into two distinct groups: mandatory which consist of non-negotiable criteria and non-mandatory criteria which consist of value added criteria.

The primary contribution of this paper is clarifying list of general criterions to assess and select the best ITO vendor to deliver the expected IT services. With a clear understanding on each of the criterions, it is hoped that a client organization could have better and richer information to properly consider which vendor to be engaged.

7. ACKNOWLEDGEMENT

Thanks to Peter B Seddon from the University of Melbourne for sharing many knowledge on IT Outsourcing practice.

8. REFERENCES

[1] Ang, S., & Straub, D.W., 1998, 'Production and Transaction Economies and IS Outsourcing: A Study of the U.S. Banking Industry', *MIS Quaterly*, Vol. 22, No. 4, December 1998, pp. 535-552.

[2] Berry, J., 2006, 'So many choices', *Offshoring opportunities: strategies and tactics for global competitiveness*, John Wiley & Sons, Inc., New Jersey, pp. 111-141.

[3] Brown, D., & Wilson, S., 2005, 'The Black Book of Outsourcing: How to Manage the Changes, Challenges, and Opportunities', John Wiley & Sons, Inc., New Jersey.

[4] Cullen, S., Seddon, P.B., & Willcocks, L.P., 2005, 'IT Outsourcing Configuration: Research into Defining and Designing Outsourcing Arrangements', *Journal of Strategic Information System*s, Vol. 14, Issue. 4, pp. 357-387.

[5] Cullen, S., Seddon, P.B., & Willcocks, L.P., 2007a, 'IT Outsourcing Configuration: Case research into structural attributes and consequences', *Proceedings of the 15th European Conference on Information Systems*, St Gallen, Switzerland, June 2007, pp.1288-1300.

[6] Cullen, S., Seddon, P.B., & Willcocks L.P., 2007b, 'IT Outsourcing Success: A multi-dimensional, contextual perspective of outsourcing outcomes', Working Paper, Department of Information Systems, The University of Melbourne.

[7] Fisher, J., Hirschheim, R., & Jacobs, R., 2008, "Understanding the outsourcing learning curve: A longitudinal analysis of a large Australian company", *Information Systems Frontier*, Vol. 10 No. 2, April 2008, pp.165-178.

[8] Galivan, M.J., & Oh, W., 1999, 'Analyzing IT outsourcing relationships as alliances among multipleclients and vendors', *Proceedings of the 32nd Annual Hawaii International Conference*, Maui, Hawaii, USA, 1999.

[9] Goles, T., & Chin, W.W., 2005, 'Information Systems Outsourcing Relationship Factors: Detailed Conceptualization and Initial Evidence', *ACM SIGMIS Database*, Vol. 36, Issue 4, Fall 2005, pp. 47 – 67.

[10] Gonzalez, R., Gasco, J. & Llopis, J., 2005, 'Information Systems outsourcing success factors: a review and some results', *Information Management & Computer Security*, Vol. 13 No. 5, 2005, pp. 399-418.

[11] Goo, J. & Nam, K., 2007, 'Contract as a Source of Trust – Commitment in Successful IT Outsourcing Relationship: An Emprical Study', *Proceeding of the 40th Hawaii International Conference on Systems Sciences*, 2007, p. 239a.

[12] Gottschalk, P. & Solli-Sæther, H., 2005, 'Critical success factors from IT outsourcing theories: an empirical study', *Industrial Management & Data Systems*, Vol. 105 No. 6, 2005, pp. 685-702.

[13] Hirschheim, R.A., & Lacity, M.C., 2000, 'The Myths and Realities of Information Technology Insourcing', *Communications of the ACM*, Vol. 43, No. 2, February 2000, pp. 99-107.

[14] Hyder, E.B., Heston, K.M., & Paulk, M.C., 2006, 'The eSCM-SP v2.01: Model Overview', in *ITSqc: Models: eSCM-SP*, accessed 31 July 2008, from <http://itsqc.cmu.edu/models/escm-sp/index.asp>

[15] Kaiser, K.M., & Hawk, S., 2004, 'Evolution of Offshore Software Development: From Outsourcing to Cosourcing', *MIS Quaterly Executive*, Vol. 3, No. 2, June 2004, pp. 69-81.

[16] Lacity M.C., & Rottman J., 2008, 'Offshore outsourcing of IT work', *Offshore Outsourcing Of IT Work: Client and Supplier Perspectives*, Palgrave, United Kingdom, pp. 1-53.

[17] Lacity M.C., & Willcocks L.P., 1998, 'An Empirical Investigation of Information Technology Sourcing Practices: Lessons from Experience', *MIS Quaterly*, Vol. 22, No. 3, September 1998, pp. 363-408.

[18] Lacity M. C., Willcocks L. P., & Rottman J. W., 2008, 'Global outsourcing of back office services: lessons, trends, and enduring challenges', *Strategic Outsourcing: An International Journal*, Vol. 1, No. 1, pp. 13-34.

[19] Levina, N., & Ross, J.W., 2003, 'From the Vendor's Perspective: Exploring the Value Proposition in Information Technology Outsourcing', *MIS Quarterly*, Vol. 27, No. 3, September 2003, pp. 331-364.

[20] Oshri, I., Kotlarsky, J., & Willcocks, L., 2007, 'Managing Dispersed Expertise in IT Offshore Outsourcing: Lessons from Tata Consultancy Services', *MIS Quarterly Executive*, Vol. 6, No. 2, June 2007, pp. 53-65.

[21] Quinn, J.B., & Hilmer, F.G., 1994, 'Strategic Outsourcing', *Sloan Management Review*, Summer 1994, pp. 43-55.

[22] Rottman, J.W., & Lacity, M.C., 2008, 'A US Client's Learning from Outsourcing IT Work Offshore', *Information Systems Frontiers*, Vol. 10, No. 2, April 2008, pp. 259-275.

[23] Schniederjans, M.J., Schniederjans, M.S., & Schniederjans, D.G., 2005, 'Methodologies for selecting outsourcing-insourcing partners', *Outsourcing and insourcing in an international context*, M.E. Sharpe, London, pp. 125-149.

[24] Seddon, P.B., Cullen, S., & Willcocks, L.P., 2002, 'Does Domberger's Theory of "The Contracting Organization" Explain Satisfaction with IT Outsourcing?', *International Conference on Information Systems (ICIS)*, Barcelona, December 2002.

[25] Seddon, P.B., & Cullen, S., 2007, 'Configuration misfit as a determinant of problems with ICT outsourcing', Working Paper, Department of Information Systems, University of Melbourne, Melbourne, AUSTRALIA.

[26] Sparrow, E., 2003, 'Choosing a service provider', *Successful IT outsourcing: from choosing a provider to managing the project*, Springer, London, pp. 67-98.

[27] Willcocks, L.P., Cullen, S., & Lacity, M.C., 2007, 'The Outsourcing Enterprise: The CEO guide to selecting effective suppliers', *The Outsourcing Enterprise series*, accessed 11 August 2008, from <http://www.logica.com/the+outsourcing+enterprise:+the+ceo+guide+to+selecting+effective+suppliers/400009148>

[28] Willcocks, L.P. & Lacity, M.C., 2006, *Global Sourcing of Business and IT Services*, Palgrave Macmillan, New York.

Information Retrieval on MARC Metadata

Adi Wibowo
Informatics Department
Petra Christian University
Siwalankerto 121-131 Surabaya
+62312983455
adiw@peter.petra.ac.id

Rolly Intan
Informatics Department
Petra Christian University
Siwalankerto 121-131 Surabaya
+62312983455
rintan@peter.petra.ac.id

Irawan Arifin
Informatics Department
Petra Christian University
Siwalankerto 121-131 Surabaya
+62312983455

ABSTRACT
A Library is usually comprised of hardcopy collections. Hardcopy collections are usually represented by MARC metadata. MARC metadata stores title, multiple authors, subject, publisher, and several other data about the collection. To retrieve hardcopy metadata a library usually use a simple text matching algorithm. The weakness of text matching algorithm is that this algorithm cannot sort its results based on relevance therefore its usability is reduced. This study proposes the use of term expansion, vector based similarity measurement, and collections borrowing history to calculate each metadata record's relevance with user's query terms.

Keywords
Similarity measurement, term expansion, MARC metadata, library.

1. INTRODUCTION

Petra Christian University Library has two types of collections. They are hardcopy collections and softcopy (digital) collections. Hardcopy collections consist of books, journals, magazines, CDs, DVDs, Cassettes, maps, etc. Digital collections consist of theses, eDimensi, Petr@rt Gallery, Petra iPoster, Petra Chronicle, and Surabaya Memory.

Until January 2010 the numbers of hardcopy collections are 111,429 title and 139,787 exemplars. The numbers of digital collections are 12,845 titles and 111,196 resources. This research concerns only hardcopy collections.

Members of library make the search using Catalog Module of SPEKTRA (Sistem Informasi Perpustakaan Universitas Kristen Petra). SPEKTRA can be accessed at http://dewey.petra.ac.id.

Each hardcopy collection is stored using MARC metadata [1]. The algorithm used by Catalog Module to find relevant collections from MARC metadata database is text matching. Because the text matching algorithm cannot sort search results by their relevancies, a new approach is needed to search and sort search results based on their relevance with member's query terms.

Table 1 shows catalog module access log for 1 year (January – December 2009). It is apparent that catalog user only used 1 or 2 terms when they searched, i.e. 367574 (75.1%) searches. Only 121842 (24.8%) searches that use more than two terms.

With only one or two terms used at every search process, the relevance calculation is more difficult. Term expansion is needed using thesauri, conceptual fuzzy, or other term expansion methods to increase relevance level of retrieved collections to query terms.

Table 1. Number of Terms (Word) used at Each Query (January-December 2009)

Number of Word	Queries
1	190042
2	177532
3	70348
4	28754
5	12323
6	6858
7	2359
8	1200

Table 2 shows that most used query methods are by title (58.7%), followed by subject and by author at position 2 and 3 respectively. But we can see that there is large difference in value between search by title (58.7%) and search by subject (15.1%) and search by author (12.7%). Despite this large difference, search by subject percentage is still significant enough to indicate the desire of members to find a collection using the context of the collection, besides the title of the collection.

Table 3 shows that the five top terms used in the query are stop words. Obviously this will slow down the query process. To overcome this problem stop words need to be removed from queries and collections metadata.

Another problem faced by information retrieval on MARC metadata is that the number of terms that usually stored for every collection only ranges between 6-15 terms. This is due to:

- Metadata fields that contain terms that can be used by IR are: title (field 130a, 245a, 245b, 490a), author (field 100-111), subject (field 600), and classification (field 099a). The number of terms in each of MARC field varies between 0-8.
- Not every collection have subject. Collections that do not have subject amounted to 27,295 (30.9%). Collections that have only one subject amounted to 32,541 (36.8%). Collections that have two subjects amounted to 20,320 (23%) of the total collection of 88,353.

Table 2. Most Used Query Methods (January-December 2009)

Collection	Method	Source	Queries
Hardcopy	Simple	Title	559080 (58.7%)
Hardcopy	Simple	Subject	144093 (15.1%)
Hardcopy	Simple	Author	120692 (12.7%)
Digital_thesis	Simple	Title	71269 (7.5%)
Hardcopy	Advanced	Title	33486 (3.5%)
Digital	Simple	Title	23563 (2.5%)

Table 3. Most Used Query Terms (January-December 2009)

Term	Count
Of	16060
And	10927
The	9531
Dan	9028
In	6186
Manajemen	5883
Management	5362
Komunikasi	4629
Indonesia	4244

The challenge that has to be answered by IR algorithm is how to find collections that relevant to the query with small number of terms that represent the collections.

2. RELATED WORKS

Until now the search process on metadata database typically rely on text-matching search. The disadvantage of this system is that this algorithm does not produce relevancy ranking of search results. Users are forced to examine all search results before obtaining the desired collections.

The approaches to search metadata database are usually to use metadata database and full text, as suggested by Tereza Iofciu and Christian Kohlschutter [2], and David Wood [3]. But there is still no algorithm that relies on metadata only

3. OVERVIEW

When a catalog user entered query terms, a retrieval process is started by conducting query term expansion using Jaccard Coefficient. Jaccard Coefficient is used to find similar terms for every term entered by the user. This process is needed to overcome small number of terms initially entered by a user. Jaccard's Coefficient will use a controlled vocabulary that is created from hardcopy collections' MARC metadata itself. Therefore the controlled vocabulary will be independent from language used by metadata, and reflect the relation between terms in hardcopy collections.

After getting a set of terms from query expansion, those term set is used to determine the most similar metadata records from a query using the Cosine Similarity Measurement. Cosine Similarity Measurement is an algorithm that uses vector as a representation of a document. Cosine Similarity is not affected by the length of the document (number of terms in each record of metadata).

To help determine relevancy ranking, the same principle from Sergey Brin and Larry Page is used. Brin and Page suggest that a popular web page, which is linked by a lot of other web pages, has higher people's subjective idea of importance [4]. By using this similar principle, this research suggests that collections' importance level also determined by how many that collection has been borrowed. But it is also need to be noted that the total borrowing of collections depends also on how long it has been owned by the library, and be allowed to be borrowed by a user. So the number of collection borrowing needs to be normalized.

Final score of collection D is calculated from the sum of relevance values obtained from the cosine similarity method, and the normalized value of the amount of borrowing P.

4. SYSTEM DESIGN

There are several phases to attain relevance value for a collection Di to a query.

1. Remove stop words from (a) collections' metadata records and (b) from user's query terms.
2. Perform query term expansion,
3. Determine similarity between a query and collection's metadata record,
4. Determine normalized borrowing value, and
5. Determine final score for a collection.

First phase (1a) and fourth can be performed whenever there is a change in the metadata to decrease execution time. While second, third, and fifth phases must be performed when there is a query process that is initiated by a user entering query terms.

4.1 Stop word

Before every term at metadata records can be used by phase 3, stop word from those terms needs to be removed. Stop word is a list of common or general terms (e.g., prepositions, and articles) that are not significant because they appear in too many records. Examples of stop words in English are 'a', 'the', 'an', 'for', 'of', etc. Example of stop words in Indonesian Language are 'di', 'ke', 'dari', 'bahwa', 'pada', etc.

To remove stop words this research use English stop word list from Gerard Salton and Chris Buckley [5]. As for Indonesian Language,

this research uses a list from Indonesian Grammar from Moeliono [6].

4.2 Query Term Expansion

Query term expansion use metadata fields only from the collection title (field 130a, 245a, 245b, 490a). It is because other fields including subject and classification use controlled vocabulary. The use of controlled vocabulary caused many of those fields have the same contents.

To perform query term expansion each document is assumed as a multiset of terms.

Doc 1 = $\{t_1, t_2\}$

Doc 2 = $\{t_1, t_1, t_2\}$

Doc 3 = $\{t_1, t_3\}$

Doc 4 = $\{t_2, t_3, t_4\}$

Doc 5 = $\{t_3, t_5, t_3, t_5\}$

To count a weight of a term to a document, it is defined that $D=\{d_1, d_2,..., d_n\}$ is a set of n documents. $T=\{t_1, t_2,..., t_m\}$ is a set of m terms. The weight w_{ij} of term i to document j is calculated as equation 1.

$$w_{ij} = \frac{tf_{ij}}{\sum_{k=1}^{m} tf_{kj}} \quad (1)$$

By using equation 1 it can be calculated that the weight of term 1 to document 2 is:

$$w_{12} = \frac{tf(t_1, d_2)}{tf(t_1, d_2) + tf(t_2, d_2) + tf(t_3, d_2) + tf(t_4, d_2) + tf(t_5, d_2)} \quad (2)$$

Similarity between terms will be calculated using Jaccard's Coefficients. It will use min and max based on T-Norm and T-Conorm that usually used in intersection and union in fuzzy set. The similarity equation between terms is shown in equation 3.

$$Sim(t_1,t_2) = \frac{\sum_{i=1}^{n} \min(w_{1i}, w_{2i})}{\sum_{i=1}^{n} \max(w_{1i}, w_{2i})} \quad (3)$$

Only five most similar terms to each particular term will be used as an expansion.

4.3 Similarity Measurement

At the similarity measurement process, terms that originated from title and author fields have to be given more weight than terms from other fields. It is because terms from subject and classification are from a controlled vocabulary so the possibility of the same terms is very high. The terms from title and author are not from controlled vocabulary so they are more indicative of the differences among collection's metadata records.

To give more weight to title and author terms, tf for every term from title and author is multiplied by 2.

Similarity measurement is performed by using cosine similarity measure.

$$Sim(Q, D_i) = \frac{\sum_{j} W_{Q,j} W_{i,j}}{\sqrt{\sum_{j} W_{Q,j}^2} \sqrt{\sum_{i} W_{i,j}^2}} \quad (4)$$

$W_{Q,j}$ is a weight of term j from query, and $W_{i,j}$ is a weight of term j from collection i. Each weight is calculated from tf*idf value.

$$w_{i,j} = tf_{i,j} * \log\left(\frac{D}{df_j}\right) \quad (5)$$

$tf_{i,j}$ is a number of term j in a collection metadata i, df_j is a number of collections that contain term j, and D is total number of collections.

4.4 Collection Transaction History

One library collection is defined as a set of collection exemplars which have similar metadata. Similar metadata means that the set of exemplars have similar title, author, subject, and ISBN. The number of borrowing of one collection is counted as the total number of borrowing transactions of its exemplars.

To normalize the collection borrowing number, there are four things that need to be concerned.

The first is that current borrowing transactions are more important than the past borrowing transactions. It is because they will represent library member interests more accurately.

The second is that the age of each exemplar is different. There are exemplars that had been acquired by library several years ago, but there are also exemplars that were acquired just several months ago. This age differences affect the number of borrowing transactions to these collections.

To overcome these problems, normalization is done by counting exemplar's age in days, and calculates constant k based on transaction age relative to exemplar age.

$$k = \frac{\textit{number of elapsed days since acquisition to transaction}}{\textit{age of collection in number of days}} \quad (6)$$

Example: a book was acquired by a library one year ago. Borrowing transaction that is conducted one month since purchasing of the book has a constant k1 = 30/365 = 0.0822. While borrowing transaction for the same book that is conducted 6 months from purchasing of the book has a constant k2 = 180/365 = 0.4932.

The third problem is the number of exemplars that are allowed to be borrowed for each collection is different with other collections. For example, one collection can have 4 exemplars allowed to be borrowed. But another collection can only have 1 exemplar allowed to be borrowed. This can also affect the possible total number of transactions for each collection. To avoid this problem the sum of constant k for each exemplar of collection D_i is divided by the

number of exemplars that are allowed to be borrowed from the collection, N.

$$P(D_i) = \frac{\sum k_i^{e1}}{N^{e2}} \quad (7)$$

To normalize P(D) then after P(D) for each collection is calculated, each score needs to be divided by the biggest P(D).

The fourth problem is that not every collection has borrowing transaction history. Collection's exemplar that categorized as tandon, reserved, theses, and reference usually are not allowed to be borrowed. They are allowed to be read only at the library. Exemplars that are allowed to be borrowed are text book, fiction, popular, etc. One collection can consist of several reserved exemplar, and also several text books. Other collection can only have tandon exemplars. So automatically this collection has no borrowing transaction history. Also new collections have no history of borrowing transaction.

To overcome this problem, two conditions of collections need to be identified. First condition is collections that part or all of their exemplars can be borrowed. For this first condition, equation (4) will be used. Second condition is collections that the entire exemplars cannot be borrowed, including the new collections. Collection is called new if its age is less than 6 months, and still has no history of borrowing transaction. For the second condition, $P(D_i)$ will be automatically given score 0,5.

$$P(D_i) = \begin{cases} \frac{\sum k_i^{e1}}{N^{e2}} & if\ condition\ 1 \\ 0{,}5 & if\ condition\ 2 \end{cases} \quad (8)$$

4.5 Final score

To calculate final score that determine collection's rank, similarity value is added to normalize borrowing score.

$$R(D_i) = aSim(Q, D_i) + bP(D_i) \quad (9)$$

$R(D_i)$ is total score of each collection D_i. *a* and *b* are coefficients used to determine the influence of each score component to final score. *a* and *b* ranges between zero and 1 ($a + b = 1$).

5. EVALUATION

To determine proper e1 and e2 several tests need to be performed. Using 'orientasi siswa' and 'corporate' as test keywords, the tests ran several times with e1 between -0.1 and -0.9, and e2 between 0.1 and 0.9. For each test iteration top-10 collections was retrieved. First position collection was given a score 10, second position collection was given 9, and so on. All results were given to participants to be rated. Each participants would rate each collection with score 1 as not accurate, and 4 as accurate. Score for each position was multiplied with score from participants. The results of test are shown at table 4 and table 5. The biggest average score achieved when e1=-0.9, and e2=0.9.

Table 4. Results of Tests to Find Proper e1

e1	Average Score	
	'orientasi siswa'	'corporate'
-0.1	112.2	127.6
-0.3	112.2	129
-0.5	112.2	129
-0.7	112,8	129
-0.9	112,8	129.6

Table 5. Results of Tests to Find Proper e2

e1	Average Score	
	'orientasi siswa'	'corporate'
0.1	112.2	128.4
0.3	112.2	128.6
0.5	1112.2	128.6
0.7	112.2	128.4
0.9	112.8	133.6

The similar tests also performed to determine proper value of a. It is found that the biggest average score achieved if a=0.8.

Compared with the old system, this algorithm can expand the query results. When use 'makan' (english = 'eat') as keyword, this algorithm can also find collections that also use word 'minum' (english = 'drink') and restoran (english = 'restaurant').

6. CONCLUSION

This study suggests an alternative search algorithm for collection's metadata which is traditionally used by the library. This study suggests using keyword expansion to every query keywords given by users since the number of words in queries is usually very small. To help the ranking of retrieval result, borrowing transaction history can be used to help distinguish the collection using current member's interest trend.

7. REFERENCES

[1] ---, *MARC 21*, http://lcweb.loc.gov/marc/

[2] Teresa Iocifu et al, *Keywords and RDF Fragments: Integrating Metadata and Full-Text Search in Beagle++*, L3S Research Center – University of Hanover, 2005
[3] David Wood, *Metadata Searches for Unstructured Textual Content*, Tucana, 2002
[4] Sergey Brin dan Larry Page, *The Anatomy of A Large-Scale Web Hypertextual Search Engine*, Computer Science Department – Stanford University
[5] Chris Bukley dan Gerald Salton, *Stopword List*, Cornell University
[6] Moeliono, A.M. et.al. *Indonesian Grammar*. Balai Pustaka: Department of Education and Cultures, 1988.

Learning Management Systems' Integration

N. S. Linawati
Telecommunication System,
Dept. of Electrical Engineering
Udayana University
Bali, INDONESIA
Phone/Fax.: +62 361 703315
linawati@unud.ac.id

Putra Sastra
Telecommunication System,
Dept. of Electrical Engineering
Udayana University
Bali, INDONESIA
Phone/Fax.: +62 361 703315
putra.sastra@unud.ac.id

P.K. Sudiarta
Telecommunication System,
Dept. of Electrical Engineering
Udayana University
Bali, INDONESIA
Phone/Fax.: +62 361 703315
sudiarta@unud.ac.id

ABSTRACT

LMSs implementation in education institutions either in distance learning, e-learning, or in blended learning has increased. Moodle is a common open-source LMS that is applied in many education institutions. However some universities have applied e-learning on more than one LMS platforms such as Wordpress and Moodle. Therefore to increase the application of Moodle as an e-learning system, it should be an integration or synchronization between Wordpress and Moodle. This can assist teachers who have a lot of learning material in their blog to migrate to Moodle. Hence the paper proposed two-ways synchronization interface application for both Moodle and Wordpress. The aim is to improve teachers' response in utilization of Moodle in their classes.

The results show satisfied outcomes. The interface can present learning contents from Wordpress into Moodle or vice versa easily. It is secure as the teachers and student have to login using student/teacher ID, course ID, username and password

Keywords

LMS, Interface, Synchronization.

1. INTRODUCTION

Mostly higher education institutions have been improving their education quality through ICT employment. There are many kinds of ICT applications for learning process such as computer-based learning, blended learning, distance learning and LMS (Learning Management System) or CMS (Content Management System). Many kinds of open source LMSs or CMSs to be chosen such as Moodle, Blackboard, Wordpress, and Jombla. Some universities have put their learning contents on Web based CMS and some have developed using pure LMS like Moodle. In Udayana University, most of teachers have put their learning material on their personal blog http://staff.unud.ac.id/~teacher_name that using Wordpress and few of them have utilized Moodle that is placed in server http://belajar.unud.ac.id. In order to increase the application of Moodle as an e-learning system in Udayana University, it should be an integration or synchronization between Wordpress and Moodle (Modular Object-Oriented Dynamic Learning Environment). This can assist teachers who have a lot of learning material in their blog to migrate to Moodle. They only open Moodle by clicking http://belajar.unud.ac.id, and their materials are automatically inserted into Moodle. Hence the paper proposed two-ways synchronization interface application for both Moodle and Wordpress. The aim is to improve teachers' response in utilization of Moodle in their classes.

2. LEARNING MANAGEMENT SYSTEM AND E-LEARNING

General definition of e-learning is the delivery of content via all electronic media, including the Internet, intranets, extranets, satellite broadcast, audio/video tape, interactive TV, and CD-ROM. Thus it is possible to be implemented for both face-to-face meeting and distant learning. For that reason, today the involvement of the use of technology to deliver training material to a target audience in a cost-effective, productive and sustainable manner is essential especially in education system. However the implementation of the learning systems requires high creativity that is time consuming. This applies to both the system and the educational content.

Some papers report that technology application in learning process yielded better students' performance. The results were the increase of the students' scores, most of the students expressed their interest to the subject, their understanding of the subject benefit in science and engineering, and their approval of utilizing learning media [1]. Furthermore is a report from U.S. Department of Education [2]. The report confirmed that on average, students in online learning conditions performed better than those receiving face-to-face instruction. The difference between student outcomes for online and face-to-face classes—measured as the difference between treatment and control means, divided by the pooled standard deviation—was larger in those studies contrasting conditions that blended elements of online and face-to-face instruction with conditions taught entirely face-to-face. The effects of competitive learning on the satisfaction and the academic achievement of telecommunications students were examined in [3]. The paper presented significant results on the use of competitive e-learning tools in university students' outcomes and satisfaction.

There are several factors that need to be examined before adopting an e-learning solution. A comprehensive E-learning solution comprises three key elements has been proposed in [4], as shown in Figure 1.

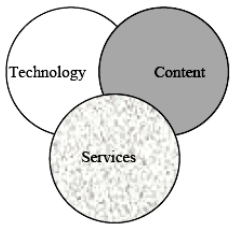


Figure 1. Comprehensive E-learning solution [4]

Technology, content development, and services should be considered to achieve a viable and sustainable e-learning system. Technology plays a fundamental role in facilitating e-learning by allowing for a range on content delivery options. Figure 2 illustrates the range of technologies that have positioned E-learning as a viable institution training option.

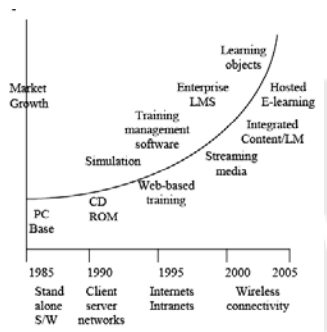


Figure 2. E-learning Technologies [5]

The global economic expansion, networking advances, together with the expansion of the Internet and the emergence of e-learning encouraged the development of LMS products. LMS provided the means for human-resource managers to manage both classroom training and the growing body of E-learning content. LMS is a software product that automates the administration of training events [6]. The LMS registers users, tracks courses in a catalogue, and records progress from learners; it also provides reports to management. An LMS is typically designed to handle courses by multiple publishers and content providers. LMS is a web-technology based, such as WebCT, Moodle, etc., conferencing and discussion systems, and rich multi-media content.

Moodle is one of the LMS products. Its application is getting popular in education institutions as it is open-source products under GNU license. Moodle is packet software that is consists of MOODLE, Apache, MySQL and PHD applications [7]. MOODLE has many advantages [8] for examples it is suitable for any kind of learning process such as online learning, blended learning, and distance learning; it can support more than 1000 courses; it has reliable security; it supports more than 45 languages; and it has three management utilizations, i.e. site management, user management, and course management.

Other proprietary LMS product named adaptive hypermedia courseware (AHyCo) has been implemented in a blended e-learning model [9]. The model is based on a mixture of collaborative learning, problem-based learning (PBL) and independent learning, in a course in Information Science, at the University of Rijeka, Croatia. The results showed that students were satisfied with the pedagogical approach, and their academic achievements were also better than expected. Particularly important is that the dropout rate was greatly diminished, which could be related to students' satisfaction with the support they received from the instructor and the system.

3. IMPLEMENTATION PHASE

Implementation phase consists of two stages, i.e. a preparation and development stages. In this phase, Moodle was proposed to be implemented in e-learning system in electrical engineering department, Udayana University. The full package application was installed in local server and acted as online LMS. Then Wordpress was installed in web server as teachers' personal blogs.

3.1 Preparation Stage

Many works should be accomplished in preparation stage. Details of works were preparing hardware and software, exploring all features and plug-ins of Moodle and Wordpress, and designing testing tools. Server, personal computers, and notebooks are major hardwares in this stage. Moodle as the main application was download from http://www.moodle.org and run in Linux environment. Then two-ways synchronization interface for both Moodle and Wordpress was designed. The interface required supporting from other applications, i.e. php, MySQL, SSH, Apache, HTML, and Java script.

3.2 Development Stage

In this stage, two-ways synchronization interface was designed and developed. Synchronization only can be done by authenticated users for example authentication of course id, student id, username and password. Therefore the system is secured. Moreover the interface will help teachers to insert their learning materials automatically from Wordpress into Moodle by simply clicking interface button. Figure 3 describes the synchronization process between Moodle and Wordpress.

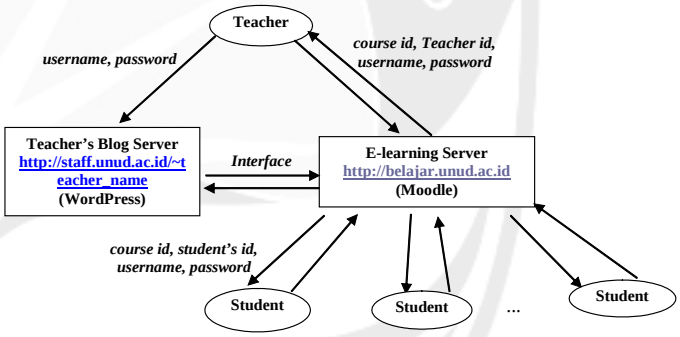


Figure 3. The synchronization Design

4. RESULTS

A two-ways synchronization interface for Wordpress and Moodle yields good results. The synchronization has good security as it is implemented only by authenticated users. Figures 5 and 6 present the interface implementations.

4.1 Interface of Wordpress to Moodle

Below is steps to display personal blog used Wordpress into e-learning system used Moodle:

- Address of RSS (Rich Site Summary) of the blog (WordPress) has to be identified, for example http://www.unud.ac.id/eng/?feed=rss2. Moodle in default mode has no RSS service of any sites, therefore a plug-in must be added. *Newsfeed* is the plug-in to be added to display RSS of a site.
- Add *newsfeed* plug-in
- Place *newsfeed* file in blocks directory of Moodle

- Login in Moodle as an administrator, then choose notifications.
- If there is no error occurred, choose continue.
- Then administrator has to set RSS feed as follows. In administrator page, choose modules, then blocks and remote rss feeds. There are three options as seen in Fig. 4, i.e. '*Entries per feed*' is for number of RSS, '*Timeout*' is for decision of RSS feed expired, and '*Submitters*' is for someone who has authorization to add or edit RSS feed.
- Click add/edit feeds, followed by RSS address of site and click Add, then Save Changes.
- '*Editing on*' menu for adding or editing the RSS feed.

Figure 4. RSS Feeds configuration in Moodle

Figure 5. Blog (WordPress) displays in Moodle

4.2 Interface of Moodle to Wordpress

On the other hand, learning contents in Moodle can be displayed in Wordpress using RSS too, as presented in Figure. 6. However RSS of Moodle has to be created in Moodle itself as follows:

- Copy file *block_recent_activity.php*, *config_instance.html*, and *rsslib.php* into directory */blocks/recent_activity/*.
- Copy file *file.php* into directory */rss*.
- Login in Moodle as a teacher in the course page which the RSS will be presented.
- Click turn editing on button, then at blocks, choose *recent activity*.
- Set on *recent activity* by activating enable *rss* feeds.
- Display RSS Moodle in Wordpress by login as an administrator in Wordpress, add pages, followed by copying script below into Wordpress page (HTML).

<iframe name="Menampilkan moodle ke dalam wordpress" src="http://172.16.40.113/moodle/rss/file.php/3/3/block_recent_activity/46/rss.xml" marginwidth="0" marginheight="0" readonly ="false" vspace="0" hspace="0" allowtransparency="true" scrolling="yes" width="800" frameborder="0" height="600"></iframe>

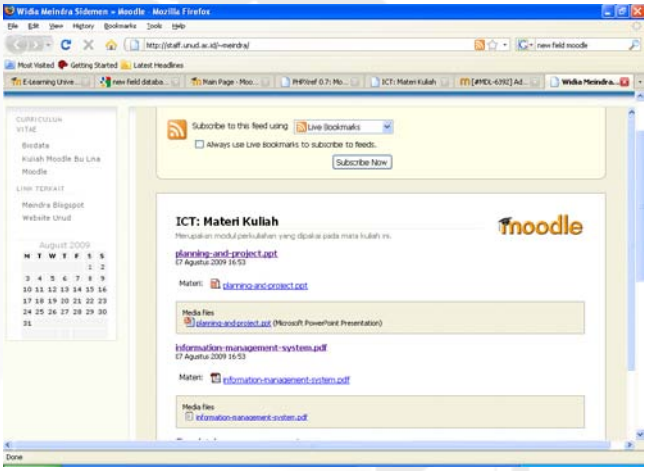


Figure 6. Moodle displays in Wordpress page

5. CONCLUSIONS

The aim of this research is to enhance LMS implementation for e-learning by developing a synchronization interface for both Moodle and Wordpress. The interface has worked well to assist teachers using LMS, and it has high security system, since course id, user'id, user name, and password are required to login into the system. Thus the interface is expected to increase the efficiency of teaching and at the same time is effective for learning. Therefore, they can take advantage of the synchronization interfaces.

6. ACKNOWLEDGMENTS

This publication is a result of research work that is funded by Indonesia Directorate General Higher Education. The authors also thank to Dr.Ir. Achmad Affandi, DEA (ITS – Surabaya) and Prof. Dr. Tsuyoshi Usagawa (Kumamoto Univ. Japan) for their valuable advices and feedbacks.

7. REFERENCES

[1] Linawati and D.M. Wiharta, 2007. E-Learning: Multimedia Application on Digital Signal Processing. In the Proceedings of the International Symposium on Open, Distance, and E-learning. (Bali, November 13 – 15, 2007).

[2] U.S. Department of Education, Office of Planning, Evaluation, and Policy Development, 2009. Evaluation of Evidence-Based Practices in Online Learning: A Meta-Analysis and Review of Online Learning Studies. Washington, D.C. DOI= www.ed.gov/about/offices/list/opepd/ppss/reports.html

[3] Regueras, L.M. Verdu, E. Munoz, M.F. Perez, M.A. de Castro, J.P. Verdu, M.J, "Effects of Competitive E-Learning Tools on Higher Education Students: A Case Study", IEEE Transactions on Education, Volume: 52, Issue: 2, P. 279-285, May 2009.

[4] Henry, P., "E-learning Technology, Content and Services", Education & Training, Vol. 43, 2001.

[5] Butler Group,"Cultural and Financial Implications of an E-learning Approach", Butler Group's Intelligence Journal, April 2002.

[6] American society for training and development (ASTD) glossary of terms. DOI= http://www.learningcircuits.org/glossary.html

[7] http://download.moodle.org/download.php/windows/MoodleWindowsInstaller-latest-17.zip

[8] http://www.moodle.org/

[9] Hoic-Bozic, N. Mornar, V. Boticki, I., "A Blended Learning Approach to Course Design and Implementation", IEEE Transaction on Education, Volume: 52, Issue: 1 on page(s): 19-30, Feb. 2009.

Mining Sequential Pattern on Sequential Data of Paint Sales Transaction Flow

Agustinus Noertjahyana
Informatics Department
Petra Christian University
agust@peter.petra.ac.id

Gregorius Satia Budhi
Informatics Department
Petra Christian University
greg@peter.petra.ac.id

Henny Kusumawati Wibowo
Informatics Department
Universitas Kristen Petra

ABSTRACT

Nowadays, information holds an important issue on every aspect in life. Business, daily life, and education sectors need information. The information discussion in this paper is more focused on the sequential data of the paint sales process to get association pattern between items bought according to the salesman's name or at some periods of time.

Analysis about the design of this application has been done from the company's data which is the sales transaction from the customer at a period of time. The data is transformed to form that can be processed by the software. After that, the data was input to a database and processed according to the mining sequential pattern algorithm. As the result are association rules and sequential association rules from the paint buying process.

This application uses Borland Delphi 7.0 software and Microsoft Office Access 2003 for the database. The data mining of this data enable the software to give information for the company about the relation between paint items that have been bought same time or sequential. Software also displays rules, graphic and tree diagram in text, that make it easier for the user to do further analysis.

Keywords

Data Mining, Association Rules, Mining Sequential Pattern.

1. INTRODUCTION

Today, many company have already used computer system as data storage transaction recording, and reporting. Data processing in small scale can be done by using simple database or spreadsheet e.g Microsoft Excel. Report which is created from those application is enough for analysing market for decision making. However, for big company which sell products in large scale, which is composed of hundred or thousand kind of product and selling type, those application are hardly to manage. There can be a missing knowledge from those data, which is significant for decisiong making, for example the pattern of the customer's purchasing.

For examples, the customer purchases which are handle hundreds of monthly sales paint. To conduct an analysis of the database, when only using the manual system, the results obtained will not be effective because such large volumes of data processed. For that, it needs a system that can provide information to users quickly and precisely. This research aims to assist decision making process by using data processing system supported by data mining, sequential pattern mining methods. Sequential pattern works by identifying or analyzing all the sequences that often appears on an item (certain paint) purchased by the customer.

With the data mining of sequential data on the purchase of paint, it will produce knowledge for paint sales. Knowledge can be useful for companies to obtain information on any paint if purchased simultaneously and paint what will be purchased in a sequence so that it can generate relationships among items as well as how much paint is purchased in a sequence that in fact different.

2. THEORY

Basically data mining is closely related to data analysis and use of software to find patterns and similarities in data collection. Retrieve valuable information which is totally unexpected to extract patterns is an unseen pattern. Progress in data collection and storage technologies quickly, enabling the organization to collect vast amounts of data. Tools and traditional techniques of data analysis can not be used to extract information from very large data, for it required a new method that can answer those needs. Data mining is a technology that combines traditional analysis methods with a specific algorithm for processing large volumes of data. "Data mining is a process to find interesting knowledge from large amounts of data stored in databases, data warehouses, or other storage media." (Han & Kamber, 2001)
Data mining analyze the data and can find important information about patterns of data, which can provide a major contribution to business strategy, knowledge base, and research and medical research. In a simple data mining or information extraction is an important or interesting pattern from existing data in large databases.
Analysis, data mining techniques in general can be orientated to existing data in a large number, with the goal of data mining can produce decisions and conclusions are guaranteed for accuracy.
The main architecture of a data mining system, in general contain the following elements:

- *Database*, data warehouse, or storage: the media in the form of databases, data warehouses, spreadsheets, or other types of storage information. Data cleaning and data integration can be performed on the data.
- *Database* or data *warehouse server*: database or data warehouse server is responsible for providing relevant data on request from the user's user data mining.
- *Data mining engine*: part of the software that runs the program based on existing algorithms.

- *Pattern evaluation module*: part of the software that find a pattern or patterns in the database. Therefore the data mining process can find the appropriate knowledge.
- *Graphical user interface*: means between the user and system for data mining communication, where users can interact with the system through a data mining query. This means provide information that can assist in the search for knowledge. Furthermore, this section allows users to browse the database and data warehouse, to evaluate the pattern that has been generated, and display patterns with different views (rules and graphics).

A proper data mining system should be built with a good algorithm, structured, fast, and can handle large amounts of data. So, when the user is dealing with a large or small database, its running time will grow proportionally. By performing data mining, interesting knowledge, high levels can be in extracting information from a database or displayed from various viewpoints. Data mining in general can be done against all sorts of data stored either in relational databases, data warehouses, transactional databases, and it was likely in a database system on the internet, such as mining, online transaction information.

2.1 SEQUENTIAL PATTERN

Data mining model based on one of two types of supervised and unsupervised learning. Supervised learning function is used to predict a value (NaiveBayes for classification). Unsupervised learning function is used to find the intrinsic structure, relations in the data that does not require a class or label prior to the learning process (a priori association rules, clustering, sequential patterns).

Sequential Pattern is a pattern that describes the time sequence of events (Han & Kamber, 2006). These patterns can be found if the data stored is relatively large, and the same object in a relatively large amount to do some action repeatedly.

The problem in data mining is finding sequential patterns described previously. Input data is a set of sequences (data-sequences). Each sequential data is a list of transactions, where each transaction is a set of items. Generally, each transaction is associated with transaction time. A sequential-pattern also consists of a list of collection items. The problem that occurs is a frequent user wants to find a sequential pattern with minimum support and a period (specific period) is determined self, where the support of a sequential pattern is the percentage of data-sequences that contain a certain pattern.
For example, the identity of customers that have registered, shopping transactions repeatedly at a store or shopping center, each data-sequence may correspond to all the choices of paint from a customer, and each transaction to the paints chosen by customer in a single order.

2.2 Generalized Sequential Pattern (GSP) Algorithm

The basic structure of the GSP algorithm is to find sequential patterns. GSP algorithm is doing multiple passes through the data. The first phase determines the support of each item, which is the number of data-sequences that include these items. At the end of the first phase, this algorithm will find out or get the item which will be frequent, which fulfill the minimum support. Each item produced a first frequent sequence consisting of the item. Each sub-sequence in each phase was originally started by a group of prospective candidate: a frequent-sequence is found or produced in the previous phase. Candidate set of candidates is used to generate new potentially frequent sequences, called candidate sequences. Each candidate-sequence has more than one item instead of the candidate sequences, so that all the candidate sequences in a phase will have items with the same number. Support from the candidate sequences were found during the process through which data exist. At the end of the phase, the algorithm will generate candidate sequences which are included in the frequent, in which the candidate is a candidate frequent candidate for the next phase. The algorithm ends when no more frequent sequences found at the end of a phase, or when no longer candidate sequences generated. There are 2 major steps in this algorithm, the candidate generation and support counting candidates.

3. SYSTEM ANALYSIS

This company of this research's object has already use computrized system to record daily transaction from the customers. The company has forty three (43) branch/warehouse in Indonesia, which are at Padang, Bekasi, Semarang, Jember, Pekanbaru, Sukabumi, Magelang, Pamekasan, Batam, Purwakarta, Solo, Makassar, Palembang, Bandung, Kudus, Kendari, Jambi, Tasikmalaya, Jogjakarta, Manado, Lampung, Cirebon, Tuban, Gorontalo, Banjarmasin, Majalengka, Surabaya, Pare-Pare, Samarinda, Tegal, Sidoarjo, Palu, Serang, Purwokerto, Kediri, Kupang, Tangerang, Pekalongan, Madiun, Ambon, Bogor, Purworejo, and Probolinggo. There are two branch will be opening soon at Sampit and Denpasar. Customer who order the products, will be received their goods from the nearest warehouse based on the request of the headquarter or salesman. Every process in and out of goods and also the process of moving goods between warehouses are logged into the computer system storage. The central computer system and warehouse are not directly connected.

The company is selling the goods through a sales division. This division will sell products to areas within the distribution area of each branch. If customers want to order the products, it can be done by contacting the salesman in request (usually the sales have been set for certain areas) or by contacting the company headquarter by telephone. Sales transactions can be done by cash or credit (28 days). Each sales transaction will be recorded in the sales invoice and submitted to the administration by entering data into the sales system.

With so many customer orders to be analyzed, the company needs the decision support system based on transaction data. The decision making process is undertaken by the branch manager and head of administration. Based on existing reports, experiences, and observations that have been done, and with many enterprise business activities makes the decision making process becomes difficult. Currently all companies analysis is still done manually.

3.1 PROBLEM ANALYSIS

The company has selling information system, purchasing information system, and stock information system. However, the problem of this company is the needed of decision support tools.

Everyday, a branch company may have hundred numbers of transactions, which is estimated as 2500–3000 transactions in a

month. Therefore, the reporting is displaying many data, which are difficult to analyze.

It is hard to overlook the relation between products in different abstraction view. The report provides the selling product sequentially, but it is hard to extract the detail. For example, reporting about a certain product category which is selling at the same time or sequentially. The company system can not provide a measure or value to describe the relation between products of the transaction.

In every decision making, the company considered a few factor, which are product category of the selling, who is the sales, , when the transaction occur. This decision is important to decide the focus of the company in the next period.

3.2 REQUIREMENT ANALYSIS

Based on the problem at 3.1, the company needs a system which is used to fulfill the requirements:

A system based on data mining to provide information for decision maker using transactional data.

A system used sequential pattern concept to describe the connection between products from any abstraction view. These points of view are item which is selling together and the sequential sales.

A system which is used generalized sequential pattern concept. This concept use product which are purchased together and sequentially, who are the sales and the transaction date as the factors or variables.

4. SYSTEM DESIGN

System design is developed using Data Flow Diagram (DFD) and flowchart, which is used to describes the system more detail.

4.1 DFD

Design is started by drawing DFD which is describe the whole data flow of the system.

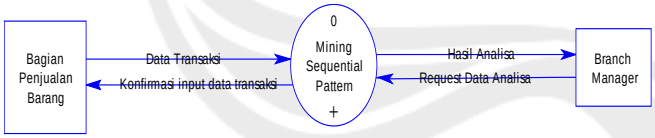


Figure 1. Context Diagram

4.2 Flowchart

Process of this application is preprocessing, generate frequent itemset, generate rules, and create report.

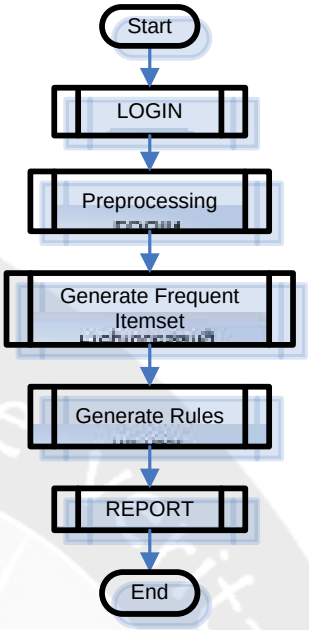


Figure 2. System Flowchart

5. RESULT

The testing stage is started from *user* authentication (login). After the login is succeed, then user is allowed to update the database by preprocessing. The sequential mining generate by setting filter and minimum support. The result can be display as rules form, graphic form, and tree form, which are followed by information needed for analyzing the mining sequential pattern.

5.1 Preprocessing Result

Preprocessing is secondary menu which is used to transform partial (by periods) or total (all) data to be analyzed. Therefore process can be shoreted. Preprocessing also used to clean data which have null value, so the process will be error-free. Updating process of prepocessing can be seen at Figure 3.

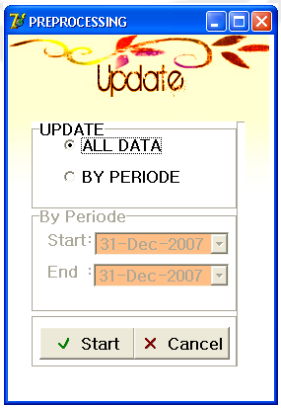


Figure 3. Update Preprocessing Menu

5.2 Analyze Result

Analyze menu is the main system of mining sequential pattern. This menu is composed of two sub menu, which are generator

menu and rule generator menu. These two menus are used to generate the possibility of frequent itemset.

5.2.1 Generator *Result*

Menu generator is the core sub-programs that work to find the pattern of purchases made by customers of paint simultaneously and sequentially in a certain period of time.

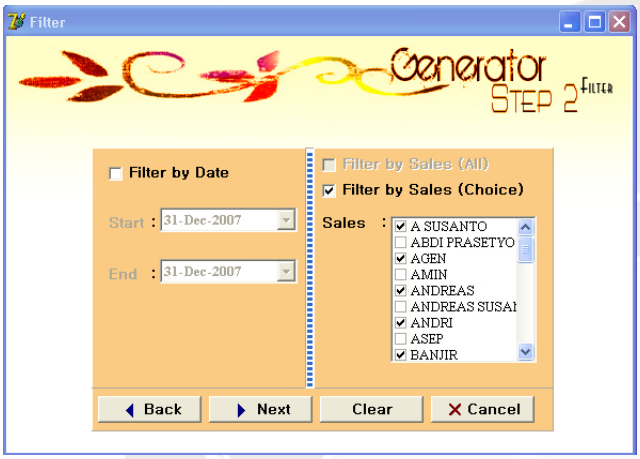


Figure 4. Generator Form

The result of generator form is frequent-dataset and candidate-sequential. Frequent-dataset can be seen at Figure 5.

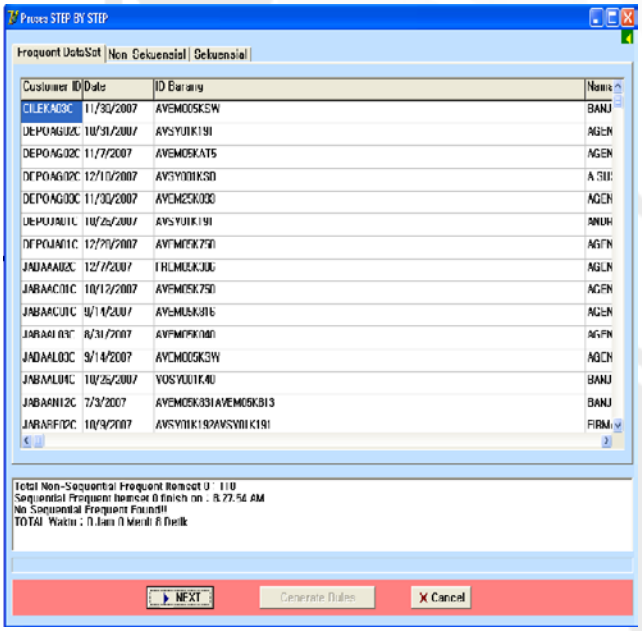


Figure 5. Frequent-dataset Form

For every candidate scanning, the system write the log file at *memolog*, *which is* showed from below of generator form and saved as text file. The content of *memolog* can be seen at Figure 6.

```
Start DateTime : 8/13/2009 8:25:44 AM
Database Use : FilterSalesDate
Total Record : 1385
Minimum Support : 4
Total Item : 385
Total Customer Id : 321
Total Non-Sequential Frequent Itemset 0 : 110
Sequential Frequent Itemset 0 finish on : 8:27:54 AM
No Sequential Frequent Found!!
TOTAL Waktu : 0 Jam 0 Menit 8 Detik
Non-Sequential Candidate 2 starts on : 8:31:58 AM
Non-Sequential Candidate 2 Finish : 8:35:01 AM
Non-Sequential Candidate 2 : 5995
Sequential Candidate 2 starts on : 8:35:01 AM
Sequential Candidate 2 finish on : 8:35:01 AM
Sequential Candidate 2 : 11990
Non-Sequential Frequent Itemset 2 starts on : 8:35:01 AM
Sequential Frequent Itemset 2 starts on : 8:36:10 AM
No Non-Sequential Frequent Found!!
Total Sequential Frequent Itemset 2 : 2
Sequential Frequent Itemset 2 Finish on : 8:36:35 AM
TOTAL Waktu : 0 Jam 4 Menit 45 Detik
Non-Sequential Candidate 3 starts on : 8:42:49 AM
Non-Sequential Candidate 3 Finish : 8:42:50 AM
Non-Sequential Candidate 3 : 108
Sequential Candidate 3 starts on : 8:42:50 AM
Sequential Candidate 3 finish on : 8:42:50 AM
Sequential Candidate 3 : 468
Non-Sequential Frequent Itemset 3 starts on : 8:42:50 AM
Sequential Frequent Itemset 3 starts on : 8:42:51 AM
No Non-Sequential Frequent Found!!
No Sequential Frequent Found!!
TOTAL Waktu : 0 Jam 4 Menit 45 Detik
No Non-Sequential Frequent Found!!
No Sequential Frequent Found!!
TOTAL Waktu : 0 Jam 4 Menit 45 Detik
Finish Time : 8:44:25 AM
```

Figure 6. Memolog File Text

5.3 Report Result

The result from report testing based on sequential and non-sequential. The result of non-sequential can be seen at Figure 7.

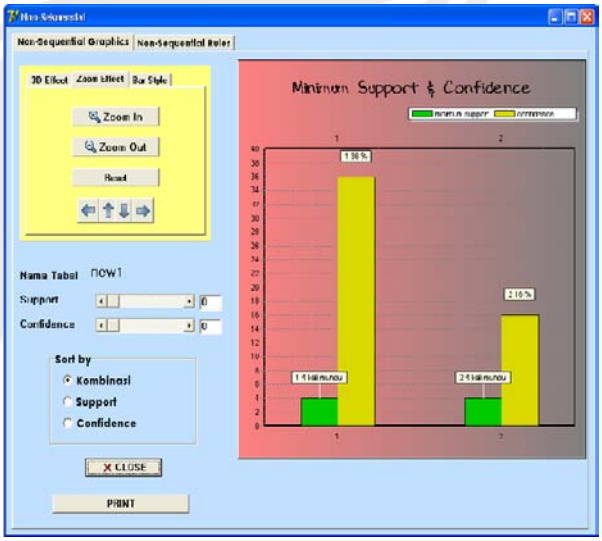


Figure 7. Result of Non-sequential

The result of sequential can be seen at Figure 8.

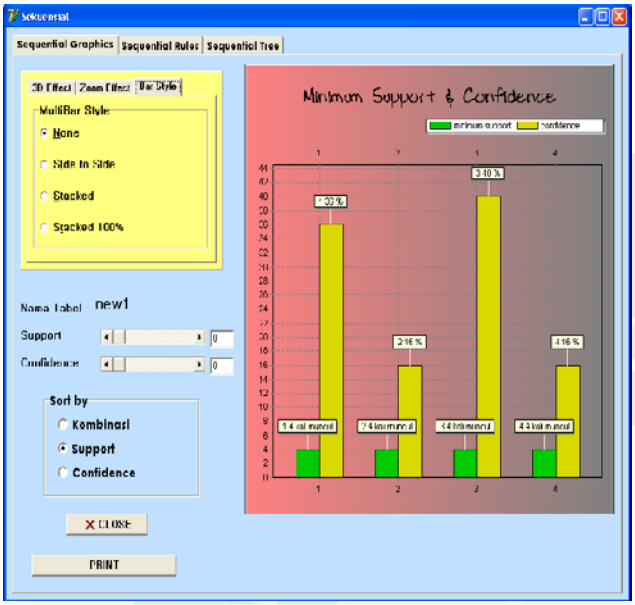


Figure 8. Result of Sequential

6. CONCLUSION

From our research, we can conclude:

- System is capable to process transactional data of paint sales for finding frequent item set which is fulfill minimum support based on items. The result is helped user to notice relation between paint which is bought from customer.
- The report result can be display as rules form, graphic form, and tree form.
- The results are determined by the minimum support, that the smaller the minimum support and confidence, resulting in programs run more slowly in a data processing rule, but can find more candidate itemset and other possibility rules.

7. REFERENCES

[1] Agrawal, Rakesh and Ramakrishnan, Srikant. (1995, March). "*Mining Sequential Patterns*". Proc. 1995 Int'l Conf. Data Eng. (ICDE '95), pp. 3-14.

[2] Fayyad, U., Piatetsky-Shapiro, G. dan Smyth, P. (1996). Retrieved Januari 11, 2007 From data mining to knowledge discovery in databases. *AI Magazine*, 37-54. http://www.aaai.org/AITopics/assets/PDF/AIMag17-03-2-article.pdf.

[3] Han, Jia Wei, and Kamber, Micheline. (2006). *Data mining: Concepts and techniques*, University of Illionis at Urbana-Champaign

[4] Han, Jiawei dan Kamber, Micheline. (2001). *Data mining: Concepts and techniques*. San Fransisco: Morgan Kaufmann.

[5] Ramakrishnan, Srikant and Agrawal, Rakesh. (1996, March). *Mining sequential patterns: Generalitazions and performance improvements*. In 5_{th} Intl. Conf. Extending Database Technology, March 1996.

[6] Ramakrishnan, Srikant and Agrawal, Rakesh. (1995, December). *Mining sequential patterns: Generalitazions and performance improvements*. Research Report RJ 1994, IBM Almaden Research Center, 650 Harry Road, San Jose, California.

Modeling School Bus for Needy Student Using Geographic Information System

Daniel Hary Prasetyo
Master Candidate Faculty of Science
University of Malaya,
50603 Kuala Lumpur, MALAYSIA
+62818586185
danielhp@perdana.um.edu.my

Jamilah Muhamad
Department of Geography
University of Malaya,
50603 Kuala Lumpur, MALAYSIA
+603-7967 5504 / 5540
jamilahmd@um.edu.my

Rosmadi Fauzi
Department of Geography
University of Malaya,
50603 Kuala Lumpur, MALAYSIA
+603-7967 5504 / 5540
rosmadifauzi@um.edu.my

ABSTRACT
Surabaya has freed the school tuition for helping the needy students. But it is not enough; they still have many children that can not go to school because the school cost is not just the tuition fee. This project tries to design free school bus for helping them in the transportation cost. Most of the design process was done in the Geographic Information System environment. It includes the data collecting, vehicle routing problem process, and the analyst process for the output. In the end, we can see that this approach can be use for defining bus routes and its characteristics.

Keywords
VRP, GIS, school bus, Network Analyst

1. INTRODUCTION
Surabaya has 492.495 school age citizen. With 270.076 at elementary school age, 114.733 at lower secondary school age, and 107.686 at upper secondary school age. While the number of school are 1.622 elementary school which consist of 564 public school and 1.058 private school, 342 secondary school which consist of 42 public school and 300 private school, and 257 high school which consist of 33 public school and 224 private school. With the number of elementary school about 5 times of secondary school and 7 times of high school, it is in general that student will travel much more distance while they moving in the higher level of study.

The number of citizen in school ages is 492.495. It divided into elementary school ages 270.076, secondary school ages 114.733, and high school ages 107.986. The government has calculated the participation rate of the education level. Participation rate indicate the comparison between the numbers of student in the certain level with the all citizen at that level ages. In early 2008 for the elementary school level the participation rate is 92.92%, in secondary school level the participation rate is 79.85%, and in the high school level the participation rate is 83.53% (Gov 2009). It mean there are more than 19 thousand citizen not in school in the elementary school ages , more than 23 thousand in secondary school ages, and more than 17 thousand in high school ages. The prime reason of the unschooled citizen is they cannot pay the schooling cost because they live in the needy family.

Surabaya city government has taken some action for helping this needy family in the education sector. The most famous action is they have freed the school cost in many public schools. In early 2008 there are 544 elementary schools and 58 secondary schools that not collect admission cost and monthly cost from their student. The rest school that still collects cost from their student can be also free for the needy student by showing the needy notes from the district government. The government not yet made some free high school because they are still focusing on the national education target, “the nine years study compulsory”. That mean, the most important is the all the Indonesian people have to study minimum in 9 years, elementary and secondary level. However, with this free admission and monthly cost, there are still a lot of children can not go to school. It because they have no extra money for buy uniform, shoes, books, and other student equipments. This project has a purpose to help them in the transportation cost.

Due to the internal limitation, we use north area, the largest number of needy student area, and it will the a scope area of all spatial and tabular data will be used in this project. Also we will use only secondary school needy student data. However the developed model will not depend on this limitation. It mean, with the same model, if the other spatial area in Surabaya have been surveyed or high school needy and places need to be included, the model will still can running well

2. LITERATURE REVIEW
Several researchers have projects related to bus routing. Some of them focus-on the use of new algorithm or advancing an existing algorithm, others are implements existing algorithm to a real world problems [1][2][3]. In advancing algorithm, Robert Bowerman and friends [10] introduce a multi-objective approach to modeling the urban school bus routing problem. Their process first groups students into clusters using a multi-objective districting algorithm and then generates a school bus route and the bus stops for each cluster using a combination of a set covering algorithm and a traveling salesman problem algorithm. A heuristic algorithm based on their formula is developed and tested with data from a sample school board location in Wellington County, Ontario, Canada.

They have defined several optimization criteria to evaluate the desirability of a particular set of school bus routes. These are :

1. Number of routes. Because the capital cost is significantly larger per bus than the incremental cost over the year, the number of routes generated should be held to a minimum.
2. Total bus route length. This criterion reduces the total length of the school bus routes.
3. Load balancing. Load balancing involves minimizing the variation in the number of students transported along each route.
4. Length balancing. This criterion involves reducing the variation in route lengths.
5. Student walking distance. This criterion balances the total distance that students walk from home to and from their bus stops against route length.

Several study have reviewed a school bus routing methodology [11][12][13][16][22][23]. Based on number of school, bus routing can be divide into many-to-one and many-to-several [19]. An Examples of many-to-one can be viewed in the work of M Fatih Demiral and friends [6] and Nayati Mohammed [5]. Both use one school location as a depot and student house location for the customer location for generating bus routes. They work with study area at Isparta, Turki and Hyderabad, India respectively. The others is Li and Fu [7] and Bektas and friends [22]. Li and Fu implement a heuristic algorithm for an existing data of a kindergarten in Hongkong whereas Bektas using integer programming for elementary school in central Ankara, Turki. They both saving 29% and 26% respectively for the generated new route compared to the current implementation. Example for the Multi School will following the others division.

Based on the location or environment of the data, bus routing can be divided into urban [8][17] and rural area [15][22]. In the urban area, the many-to-one from Bektas and friend and Li and Fu can be a good example. In the rural area, Armin Fu¨genschuh [9] take five county are in German for the student location. While the destination is multiple school, Instead of sending the bus back to the depot after having served a trip , he push to re-use the bus to serve other trips, as long as this is possible. He integrates optimization of school start times with the optimization in the school bus transportation. He believes with this integration for a single county can save up to 1 Mio. Euro – year by year. Another work in rural area focusing in the advancing algorithm used rural school data in Savigny and Forel, Swiss [14]. This two rural area environment route the buses for multiple school.

In the heuristic solution approaches for bus stop selection are classified into the location-allocation-routing (LAR) strategy or the allocation-routing-location (ARL) . The LAR strategy first determines a set of bus stops for a school and assigns students to these stops. Routes are generated for these selected stops. However, since the bus stops and the assignment of the students are determined without taking into consideration their effect on generating routes, this approach tends to generate excessive routes. In the ARL strategy, the students are allocated into clusters while satisfying vehicle capacity constraints. Subsequently, the bus stops are selected, and a route is generated for each cluster. Finally, the students in a cluster (route) are assigned to a bus stop which satisfies all the requirements given in the problem such as the maximum walking distance from home, maximum number of students that can be assigned to a bus stop, and the minimum distance apart between bus stops.

In the both concept, student assumed can walk to the bus Stop. This assumption is used in almost all of bus school routing researcher.

3. DATA COLLECTION

The readiness of data is the most important thing in the GIS project. It can consume half, and even more, of the project timeline. For this project the data are vary in the readiness , some data is ready to use and the other are still have to create. These are the data have to be collected for the VRP design process:

3.1 Sub-sub-district map

The government of Surabaya city just has a map with sub-district area in detail. That is the reason why the sub-sub-district a survey for mapping the sub-sub-district boundary must be conducted. Some area still large and not divide into smallest part. It is because those areas are not administrated by government of Surabaya city but in the control of military department, since the area are for military basis and army domiciles. The total number of sub-sub-district in this study area is 274 regions.

3.2 Needy student data

Surabaya city government has collected their needy citizen data. In 2007 there are 550.783 people in the 119.219 families detected live below the poverty line. This needy people data save their name, birthplace and birthday, address, sex, and their occupation. We query the data with age between 13 to 15 old. This data then converted in the DBF format for loaded in the ArcMap and joining with tabular data of sub-sub-district map. After this join, the amount of needy students copied in the needy population field in the sub-sub-district map in the point a. The total number of needy students spread in this sub-sub-district is 8579 students.

Beside the amount of needy student we also need a field for saving the needy density. The density field will be helpful for recounting the number of something related with a region when there are modifications in their shape and or size, for example for recounting needy in the such area in the output of clipping operation. We can construct the needy density in the specific area we want to. In this project we calculate the needy density in the 1 hectare (ha) area. If the needy density number is 4. It can say that in the 100 Meter X 100 Meter in this area we will find 4 needy students. Since the map unit is in meter, and the area of a region in the map in M2, we have to divide the area with 10.000 to get hectare before used to divide the number of needy students.

3.3 Street Map

It can be said that street map is the main object here. It need for generating the routes where the school bus will picking up the student and deliver them to the school. The government of Surabaya city is already maps their streets. But we can directly use

it for generating the bus route. First, because not all the street can be passed by the bus, we must eliminate the secondary streets.

Street map used in ArcGIS Network analyst need a special format [4]. This special requirement sometime needs an extra attention and time. We must carefully reshape each street objects with focusing in their junction. Every junction determines whether it can be use for turning to another street or not. The common street map is not separate each edge in every junction. Street map, from the government, need to be cut in almost every junction. Junctions which not being cut can be imagine that the streets across it are not in the same elevation, one street object is lying under or above the other. After separating street objects in its all junction, we must guarantee that all vertices of the streets object corresponded to these junction are perfectly patched each others. If they are not perfectly patched, when use in vehicle routing problem after build the network dataset, the bus will can not pass through that junction. For this purpose, snapping function became an important function to include in the map editor.

Some street can be passing in two ways or just one way. We also need to take attention the one way streets. This one way rule depends on the two things. First, how the sequence of the vertices in the digitizing process, and second is the value of 'ONEWAY' field . The value of oneway field is 'FT', 'TF' , or None. None mean the street is not oneway. FT stands for From-To, and TF stands for To-From. If the value is FT it's mean the street is a one way street with the direction is from the earlier digitized vertices to the latest digitized vertices. Contrary if the value is 'FT'. For example, if we want to make one way street from West to East we can take one of this two ways. First, we digitize the street along from West to East and fill the oneway field with 'FT', and the second we can digitize from East to West and fill the oneway field with 'TF'. North Surabaya city streets not have many one way streets. Most of the one way streets is as part of the double way streets.

Vehicle routing process will find the most optimal route. It may take a longer path but with more little time. For the best result the street data must be have a field that save a value about how much time is needed for passing each streets. The standard field name for this purpose is 'FT_MINUTES' and 'TF_MINUTES' which are use for saving time needed to pass the street in the same direction as the digitizing sequence, and reserve the digitizing sequence respectively. In order to filling this two field and for making the result closer to the real world, a physical survey must be conducted.

3.4 School Map

School buses will traverse the street stop near the schools. Therefore the next data we need to collect is the location of the schools. This map is already done by the government, therefore no survey and digitizing needed. There are 57 schools consist of 8 public schools and 49 private schools. The public school is a school owned by the government, and they will be get a more attention next not only as places for delivering/picking student but also for the bus depots. It is because the public schools commonly have more student capacity than private school. They also have more extent building and its surrounding area. And the most important, the government have a right for manages this areas.

3.5 Bus Depot

School buses will travel from one school to another. But before traveling in the first school in the sequence, the buses need a starting point. After visiting last school in the sequence, the buses need to end the travel in a stopping point also. This start and stop point called depot. Surabaya city have several bus depot location spreading in the city area, and they are already mapped by the government just like another public facilities. At the north area, there are two bus depot locations, one in the center and the other in the west. All of this data can be seen in figure 1 below..

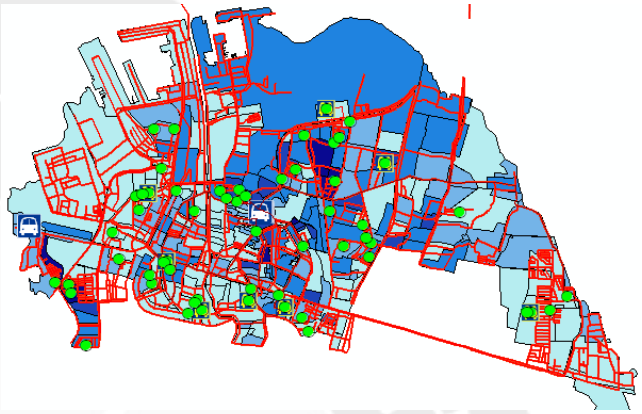


Figure 1. All needed map: sub-sub district, school, bus depot, and street map

4. VEHICLE ROUTING PROBLEM ANALYST LAYER

The route will generate by Vehicle Routing Problem tools, included in the Network Analyst of ArcGIS 9.3 because this tool will automatically generate the best route, based on Djikstra Algorithm, we just need to focus on the setting of their tool requirements.

The first step we must convert the street map into a network dataset. We can easily do this in the ArcCatalog. Since we use the standard field name like ONEWAY, FT_MINUTES, an TF_MINUTES the process will automatically detect and use the data for generate network dataset. Actually there are another field use in this process. That is the COST field. This field needed if some streets are in different cost than the others when it passed through. For example is a toll road. Since there is no toll road need to be traveled in this school bus route, we omit this field.

Another setting need to consider in this project is the turning setting. There is two types of turning used in the network analyst. First is by using turn feature and second by using the Global turn. With using turn feature, we can manage every turning rule in every junction. We can set the turn right, turn left , and u turn rule, and how much time each turn consumed . In contrast, with general turn we assume all junction have the same turning rule and time consumed. Since almost all junctions in this study area have similar characteristic, this project use the global turn for manage the turn rule. This turn will add some extra time when passed through.

After generate network dataset in ArcCatalog, the rest process will do in ArcMAP. Before network analyst can use for finding solution, we have to define a collection of setting that called Vehicle Routing Problem Class. This class will be shown in the

map as a vehicle routing problem analysis layer. Like common layer, we can then set the display property of object when shown in the map. The minimum setting for VRP class is we have to defined 3 sub-class, they are Orders, Depots, and Routes.

Orders are places that have to visit by the vehicle. It can be for delivering something, picking up something, or just a place to inspect. An order can have size or capacity. For this bus school we can set it with a needy student capacity per school. The total number of needy student is 8.581 students. Since number of school is 57 schools, the average capacity of each school for needy student is 150. We set the orders class capacity with this number, and will refine next in the chapter 4.

Another important parameter for Orders class is service time and time windows. Service time is time needed for all students doing get into or out from the bus. We set this parameter 3 minutes. Time window separated into 2 parameters, time window start and time window end which define what the start time and until when the school can be serviced as an order. School is start at 7:00 am, and we can get the students to the school 1 hour and a half before that time. Therefore these parameters we set 5:30:00 AM and 7:00:00 AM respectively.

Next class need to be set is Depots. As introduce earlier, these are to define where the bus start and end its traveling. In the school bus routing, the bus is start from bus depot, end in some place and waiting there until end of school day and then reverse its travel route back to the bus depot. For these waiting places we can use public school area, because government has an authority for using it. Therefore for the depots class we load 2 layers, bus depot layer and school layer with query to select the public school only. The location of the Depots layer is shown in Picture 3.9 below. Depot has also service time and time window parameter. Service time is set to 3 minutes and the service time is set to 5:00:00 AM to 8:00:00 AM. It's mean the depot can start loading the bus in 5:00:00 AM and has 30 minutes to travel and get passenger before visit to the first school in 5:30:00 AM.

We already have the start and stop point and places to visit while traveling. Now we are ready to define the route with some rule we want. To be considered that the less number of route mean the more school choices for the students. For example, if all the school in the study area can visit in one route, then wherever school the student want to go, that route can cover. If there are two route, than the school will divide into two path, and the student can only take to half of all school. We try first with one route, two route, three route, and stop in four route needed to cover all the schools. This four route described in the figure 2 below.

Some vehicle routing problem need to setting up the Capacities parameter. The capacities parameter is the maximum amount (for instance, volume, weight, quantity) that can be carried by the vehicle. This parameter will limit the traveling pattern when the capacity reach maximum. For school bus with one vehicle, it can be set with the total number of the seats, that about 50 seats. But, we want to discover the most optimal routes and then analyzing to reverse and get the capacities of each route, we omit this parameter with setting it with large value, in order the vehicle can carry all orders and will not limit the generated route. After setting this parameter the VRP Analyst ready to run and generate a solution. After waiting about 30 second, in intel core 2 processor, we will get the generated routes like in the figure 3 below, equipped with the orders sequence and its visited time. The red bold lines show the junction of routes. It mean one street used by two or more routes, that's also can be used for student to exchange the vehicle from one route to another.

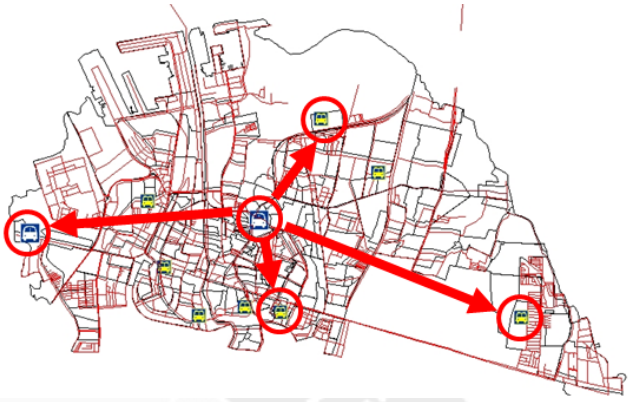


Figure 2. Start and end depot design

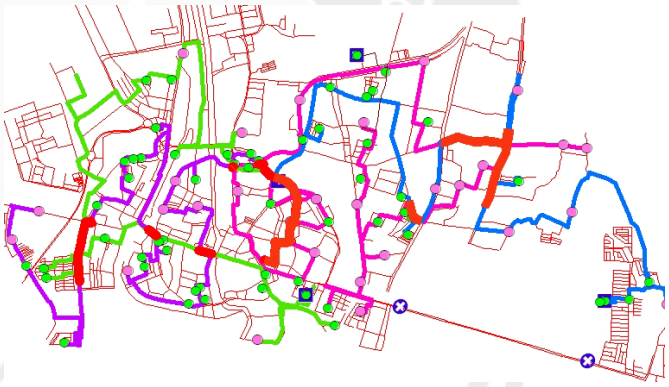


Figure 3. Routes generated by VRP analyst.

5. ROUTE ANALYSIS

There is two kinds of analyst we can get from these generated routes. We can discover covered area and the expectation of load and flow of the passenger of the bus so we can predict how many buses to be provided for getting the best service.

5.1 Covered area analyst

Now we will examine each route. Each route saves to shapefile separately for more details in analyst. After the new shapefile load in the map, we continue with creating buffer. The distance of buffer is set with acceptable walking distance for children at secondary school age from their home to the street to for ride the bus. We set this value with 200 Meter, not too far and not to close. With the buffer area, then we Clip the sub-sub-district map and get the sub-sub-district map in area with distance 200 Meter from bus route. We then save this area into new layer. This new layer still can not express the number of needy student in the area until we recalculate the wide of the area and the needy density. By this clipping and recalculating, we discovered that the west route covers 20 schools for 2677 needy students. This route start on 5:26 to 6:46, still have

about 14 minutes reserved time for others bus start after the first bus in this route. The south route covers 9 schools for 1989 needy students. It starts at 5:26 and end the journey at 6:38, more reserved time available. The north east route covers 12 schools for 2569 needy student. Start in 5:25 and end at 6:33, and become the quickest routes. Meanwhile the north route covers plenty of students, with 4317 for 20 schools. This route becomes the longest route since start at 5:26 an end at 6:55.

The above paragraph shows that the total number covered by all routes is more than the total number of needy students it is because they share their covered area. For finding this shared area, we first intersect covered area of all routes. It will give us a 'very lucky' area, the area with 4 choices route. Next we make all possible combination of intersection of three routes. Then we union all output of this combination for finding the area which can choose 3 routes or more. For finding area with exactly 3 choices route we subtract the union output with are with 4 choices route. The same steps do for finding the area with exactly 2 choices route. Make all possible combination of intersection of two routes, union all its outputs, and then subtract with the exactly 3 and exactly 4 choices route. The last, for finding the area with exactly 1 choice route, we union all routes and subtract with exactly 2 , exactly 3, and exactly 4 choices route. The four kinds area with different number of nearby route can be seen in figure 4. With recalculating all area of these steps we discover there are 4811 needy student with that just have 1 bus route nearby, 1441 students have 2 route nearby, 153 students have 3 routes nearby, and 846 students have 4 routes nearby. The total covered student is, by the sum of those numbers, 7904 or 92.12% of needy students.

Figure 4. Four kinds of needy covered area

5.2 Load and flow analyst

The covered area analyst is just show the big picture of how many people can be serviced. In this next analyst we will going closer to found how many bus needed and how its load and flow of the passenger characteristic is. We use one by one route for this analyst. In this chapter we will detailed the east route. The first step is cutting the buffer of route at the school area. It will divided the load from and to each schools. Figure 5 show this dividing east route.

Figure 5. East route divided by schools

We then calculate the amount of needy in each part. We use the same way as in the covered area analyst and continue with dissolving the output of clipping process. For this dissolve we need to get the part id, so we have to use Identify function first, before dissolving are with the same part id. After it we get parts with amount of needy in them.

Table 1. Needy load and flow for east route

No	sub-route	load	total	school capacity	rest	unused seat
1	SMP Islam Al Amal	115	115	-150	0	-35
2	SMP Islam Lil Wathon	89	89	-150	0	-61
3	SMP Muhammadiyah 16	151	151	-150	1	0
4	SMP PGRI 6	152	153	-150	3	0
5	MTs Nurul Salam	23	26	-150	0	-124
6	SMP YP 17	237	237	-150	87	0
7	SMP Muhammadiyah 15	26	113	-150	0	-37
8	SMP Tri Tunggal 7	54	54	-150	0	-96
9	SMP TarunaJaya 1	384	384	-150	234	0
10	SMP Romly Tamim	341	575	-150	425	0
11	SMP PGRI 11	45	470	-150	320	0
12	SMP Negeri 18	3	323	-150	173	0
		1620		-1800		-353

By looking at the sequence of route we take the sequence and the amount of needy in the area of each sequence into Microsoft excel, and add some calculation field like shown in the table 1 above. We assume each school allocating 150 seats for the needy student and this amount will fulfill in the bus visiting. The first trip is go from the start depot to SMP Islam Al Amal. The busses will take 115 needy student and drop all in this school. So it will make 35 seat unused. And then buses will travel again to the next schools. Sometimes the capacity number more than the picked student number, but sometimes it less. If this happen, the student will not

drop to that school but continue travel to the next school. As we see in figure table 1 , In the last school there is still 173 needy students that can not drop to school. The capacity is more than the load, but not all student can drop in to school. What we have done is just count the load from start depot to the east. We can fight this by make reserve route from east to starting depot. We assume that 173 not ride the east buses, but ride the reserve route. Because there are buses from the other way, we can use the unused seat as the capacity for the reserve route. Table 2 show that the load and flow into the table 1 split in two tables, the east route and the reserve route.

Table 2. Needy load and flow for east route+reserve route

no	Sub-route	load	total	school capacity	rest	unused seat
1	SMP Islam Al Amal	115	115	-150	0	-35
2	SMP Islam Lil Wathon	89	89	-150	0	-61
3	SMP Muhammadiyah 16	151	151	-150	1	0
4	SMP PGRI 6	152	153	-150	3	0
5	MTs Nurul Salam	23	26	-150	0	-124
6	SMP YP 17	237	237	-150	87	0
7	SMP Muhammadiyah 15	26	113	-150	0	-37
8	SMP Tri Tunggal 7	54	54	-150	0	-96
9	SMP TarunaJaya 1	300	300	-150	150	0
10	SMP Romly Tamim	300	450	-150	300	0
11	SMP PGRI 11	0	300	-150	150	0
12	SMP Negeri 18	0	150	-150	0	0
		1447		-1800		-353
	Reserve route					
1	SMP PGRI 11	3	3	0	3	0
2	SMP Romly Tamim	45	48	0	48	0
3	SMP TarunaJaya 1	41	89	0	89	0
4	SMP Tri Tunggal 7	84	173	-96	77	0
5	SMP Muhammadiyah 15	0	77	-37	40	0
6	SMP YP 17	0	40	0	40	0
7	MTs Nurul Salam	0	40	-124	0	-84
8	SMP PGRI 6	0	0	0	0	0
9	SMP Muhammadiyah 16	0	0	0	0	0
10	SMP Islam Lil Wathon	0	0	-61	0	-61
11	SMP Islam Al Amal	0	0	-35	0	-35
		173		-353		-180

With dividing the route into east route and the reserve, we can drop all the student to the school. In the east route, maximum passenger number reach in the stage 10 that travel from SMP Taruna Jaya 1 to the SMP Romly tamim with 450 passengers. Since one bus carrying about 70 students, for the east route we need 7 busses, and the reserve route 3 busses. This number is the ideal number with assumption all needy students take a bus for go to school. This number can be decreased with assuming there are amount of students travel by riding bicycle and walking to the school.

6. MODELLING THE PROCESS

Process to generated a route and follows with analysis sometime need some action that have to do repeatedly. The steps in this project also can used by another map data for another locations. A GIS programmer can make a module to reduce the steps and make it more flexible and reusable. He can follow steps in this project ttah shown in figure 9.

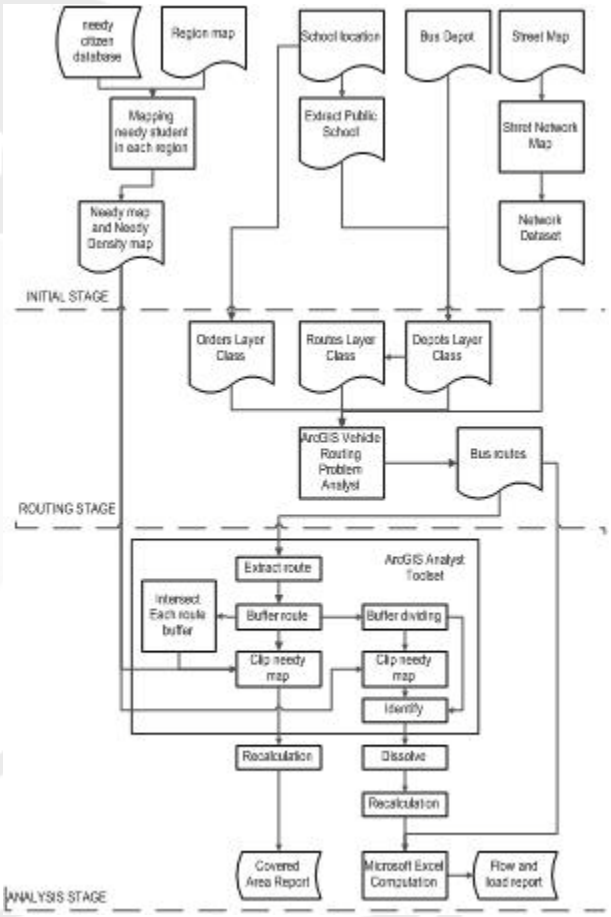


Figure 9. Model schema used in this project.

7. CONCLUSION

VRP Function, a new function in network analyst in ArcGIS 9.3 can use for finding some routes with several scenarios. For analyzing the generated routes we can compliment it with another analyst tools. For the Surabaya case study we can conclude that for ideal service in helping needy students, the government has to provide a significant number of bus schools. But with analysis model in this paper, they can choose the best distribution of bus in the limited number of buses they have.

8. REFERENCES

[1] Dorronsoro Díaz. 2007. The VRP Web. Languages and Computation Sciences department of the University of Málaga. http://neo.lcc.uma.es/radi-aeb/WebVRP.

[2] P. Toth and D. Vigo, editors.2002. The Vehicle Routing Problem. Monographs on Discrete Mathematics and Applications. SIAM, Philadelphia,PA.

[3] Bruce Golden, S. Raghavan, and Edward Wasil. 2008 The Vehicle Routing Problem: Latest Advances and New Challenges. Springer Science+Business Media, LLC 2008.

[4] ESRI Team 2008. The ArcGIS Desktop Help. ESRI Press 2008.

[5] M.Fatih Demiral, Ibrahim Gungor, Kenan Oguzhan Oruc. 2008. Optimization at Service Vehicle Routing and A Case Study of Isparta, Turkey.First Inernational Conference on Management and Economic (ICME 2008) Tirana, Albania 2

[6] Nayati Mohammed Abdul Khadir. 2008. School Bus Routing and Scheduling usng GIS. Master thesis in Geomatics, University of Gavle.

[7] L. Y. O. Li and Z. Fu. 2002. The School Bus Routing Problem: A Case Study. The Journal of the Operational Research Society, Vol. 53, No. 5 (May, 2002), pp. 552-558

[8] Lazar Spasovic, Steven Chien, Cecilia Kelnhofer-Feeley, Ya Wang, and Qiang Hu. 2001. A Methodology for Evaluating of School Bus Routing - A Case Study of Riverdale, New Jersey. Transportation Research Board 80th Annual Meeting January 7-11, 2001. Washington, D.C.

[9] Armin Fu¨genschuh,2009. Solving a school bus scheduling problem with integer programming. European Journal of Operational Research 193 (2009) 867–884.

[10] Robert Bowerman, Brent Hall, and Paul Calamai. 1995. A Multi-Objective Optimization Approach to Urban School Bus Routing Formulation and Solution Method. Elsevier Science Transportation Research Part A: Policy and Practice Volume 29, Issue 2, March 1995, Pages 107-123.

[11] Junhyuk Park, Byung-In Kim. 2009. The school bus routing problem: A review. Elsevier European Journal of Operational Research 202 (2010) 311–319.

[12] Burak Eksioglu, Arif Volkan Vural, Arnold Reisman.2009. The vehicle routing problem: A taxonomic review . Computers & Industrial Engineering 57 (2009) 1472–1483

[13] Schittekat, P., Sevaux, M., Sörensen, K., 2006. A Mathematical formulation for a school bus routing problem. Proceedings of the IEEE 2006 International Conference on Service Systems and Service Management, Troyes, France

[14] Spada, M., Bierlaire, M., Liebling, Th.M.2005. Decision-aiding methodology for the school bus routing and scheduling problem. Transportation Science 39, 477–490.

[15] Chen, D., Kallsen, H.A., Chen, H., Tseng, V. 1990, Bus routing system for rural school districts. Computers and Industrial Engineering 19, 322–325..

[16] Corberán, A., Fernández, E., Laguna, M., Martí, R. 2002. Heuristic solutions to the problem of routing school buses with multiple objectives. Journal of Operational Research Society 53, 427–435. 2002

[17] Braca, J., Bramel, J., Poser, B. and Simchi-Levi. A. 1994. Computerized Approach to The New York City School Bus Routing Problem. Technical report, Graduate School of Business, Columbia University, NY.

[18] Peter Keenan 2008. Modelling vehicle routing in GIS. Operational Research International J (2008) 8:201–218

[19] Robert A. Russell, Reece B. Morrel, Bob Haddox. 1986. Routing Special-Education School Buses. Interfaces, Vol. 16, No. 5 (Sep. - Oct., 1986), pp. 56-64

[20] Arthur J. Swersey and Wilson Ballard. 1984. Scheduling School Bus. Management Science, Vol. 30, No. 7 (Jul., 1984), pp. 844-853

[21] Ripplinger, D. 2005. Rural school vehicle routing problem. Transportation Research Record 1922 (2005), 105–110.

[22] Bektas T., Elmastas S. 2007. Solving school bus routing problems through integer programming. Journal of the Operational Research Society 58 (12)(2007), 1599–1604.

[23] Lazar Spasovic, Steven Chien, Cecilia Kelnhofer-Feeley,Ya Wang, Qiang Hu 2001. A Methodology for Evaluating of School Bus Routing - A Case Study of Riverdale, New Jersey. Transportation Research Board 80th Annual Meeting January 7-11, 2001 Washington, D.C.

Optimization SQL Server 2005 Query Using Cost Model and Statistic

Ibnu Gunawan
Informatic Engineering Department – Petra Christian University
Siwalankerto 121 – 131 Surabaya 60236 Indonesia
Telp. (031) – 2983456, Fax (031) – 8417658
ibnu@peter.petra.ac.id

ABSTRACT
MS SQL Server Query Optimizer[1] is an optimization tools that based on a cost model, the database metadata, database statistics, system resources (memory, IO, CPU) and the query itself . If we want to optimize the Query in MS SQL Server, we must optimize the factor that MS SQL Server Query Optimizer rely on. Three of that factor is cost model, statistics and the query itself. In SQL Server 2005[2] one tools that can do optimization based on statistic and query is known as DTA . These Paper will explain how the query optimizer work, then it will explain how the DTA work side by side with query optimizer and it will explain how to use the DTA to auto optimize the query.

Keywords
MS SQL Server, Database, Statistic, Query, Optimization.

1. INTRODUCTION
The query optimizer[3] in MS SQL Server 2005 selects a plan for a given query based on cost model, the database metadata, database statistics, system resources, and the query itself.

The query optimizer is the component in a database system that transform a parsed representation of an MS SQL Server 2005 query into efficient execution plan for evaluating it. Optimizer usually examine a large number of possible query plans and choose the best one in a cost-based manner.

To efficiently choose among alternative query execution plans[4], query optimizers estimate the cost of each evaluation strategy. This cost estimation needs to be accurate (since the quality of the optimizer is correlated to the quality of its cost estimations), and efficient (since it is invoked repeatedly during query optimization).

Poor performance can result due to modeling assumptions[1], outdated statistics, system resource assumptions etc. It is useful to have a tool for database engine users (DBA, application writers, and architects) and designers (developers, customer support) to understand the source of plan inefficiencies.

For example, engine designers can use it to understand why a particular join strategy resulted in poor performance whereas another known join order has better performance. Customer support can use such a tool to help narrow down the possible causes of poor performance to particular characteristics of a plan e.g. wrong join order.

In next section we describe the components of a generic query optimizer and show how statistical information can be used to improve the accuracy of cost estimations, which in turn impacts the whole optimization process.

2. BACKGROUND
The query optimizer is the component in a database system that transforms a parsed representation of an SQL query into an efficient execution plan for evaluating it. Optimizers usually examine a large number of possible query plans and choose the best one in a cost-based manner. To efficiently choose among alternative query execution plans, query optimizers *estimate* the cost of each evaluation Strategy . This cost estimation needs to be accurate (since the quality of the optimizer is correlated to the quality of its cost estimations), and efficient (since it is invoked repeatedly during query optimization). In this section we describe the components of a generic query optimizer and show how statistical information can be used to improve the accuracy of cost estimations, which in turn impacts the whole optimization process. After that we show that optimizer is part of DTA

2.1 Query Optimizer Architecture
There are several optimization frameworks in the literature [5, 6, 7, 8, 9] and most modern optimizers rely on the concepts introduced by those references. Although the implementation details
vary among different systems[4], all optimizers share the same basic structure [10], shown in Figure 1.

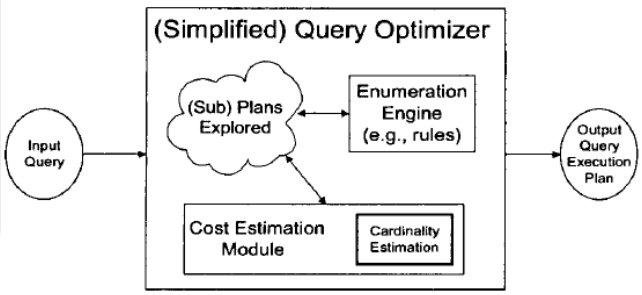


Figure 1. Simplified Optimizer's Architecture.

For each incoming query, the optimizer maintains a set of sub-plans already explored, taken from an implicit search space. An enumeration engine navigates through the search space by applying rules to the set of explored plans. Some optimizers have a fixed set

of rules to enumerate all interesting plans (e.g., System-R) while others implement extensible transformational rules to navigate through the search space (e.g., Starburst, Cascades). All systems use dynamic programming or memorization to avoid recomputing the same information during query optimization. For each discovered query plan, a component derives different properties if possible, or estimates them otherwise. Some properties (e.g., cardinality and schema information) are shared among all plans in the same equivalence class, while others (e.g., estimated execution cost and output order) are tied to a specific physical plan. Finally, once the optimizer has explored all interesting plans, it extracts the most efficient plan, which serves as the input for the execution engine.

A useful property of a query plan from an optimization perspective is the estimated execution cost, which ultimately decides which is the most efficient plan. The estimated execution cost of a plan, in turn, depends heavily on the cardinality estimates of its sub-plans. Therefore, it is fundamental for a query optimizer to rely on accurate and efficient cardinality estimation algorithms.

EXAMPLE 1. Consider the query in Figure 2(a) and suppose that IRI ~ ISl ~ ITI. If the query optimizer has knowledge that R.a < 10 is much more selective than T.b > 20 (i.e.,just a few tuples in R verofy R.a < 10 and most of the tuples in T verify T.b > 20), it should determine that plan P1 in Figure 2(b) is more eficient than theplan P2 in Figure 2(c) 3. The reason is that Pl first joins R and S producing a (hopefully) small intermediate result that is in turn joined with T. In contrast, P2 produces a large intermediate result by first joining S and T.

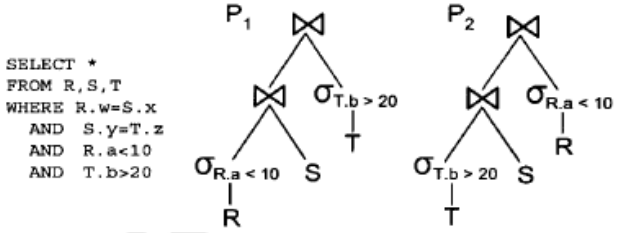


Figure 2. Query plans chosen by query optimizers depending on the cardinality of intermediate results.

Cardinality estimation uses statistical information about the data that is stored in the database system to provide estimates to the query optimizer. Histograms are the most common statistical information used in commercial database systems.

2.2 Database Tuning Advisor Architecture

DTA takes as input a workload consisting of T-SQL (SQL Server's flavor of the SQL language) statements such SELECT, INSERT, UPDATE, DELETE, stored procedure calls, dynamic SQL, and DDL statements and produces as output a T-SQL script that consists of recommendations for indexes, materialized views (called indexed views in Microsoft SQL Server), and horizontal partitioning[2].

The architecture of DTA appears in Figure 3. In Figure 3 we see that the query optimizer is a part of DTA. The details are presented in [11], and have been omitted. For clarity, we will summarize the input/output of DTA.

DTA recommends a set of indexes, materialized views, indexes on materialized views and their respective horizontal partitioning that is appropriate for the given workload. The tool relies on interfaces provided by Microsoft SQL Server's query optimizer [12, 13]. First it needs to be able to simulate partitioned tables/indexes and materialized views (referred to as hypothetical indexes and materialized views) that do not exist in the current database. Second, it needs to tell the query optimizer to optimize a query for a given hypothetical configuration (a set of partitioned tables, indexes, materialized views and indexes on materialized views).

DTA [2] searches the space of (partitioned) tables, indexes and materialized views to arrive at the best configuration for the given workload. In general, the above search problem can be prohibitively expensive. Column Group Restriction, Candidate Selection, Merging, and Enumeration (Figure 2) are individual steps in the search algorithm that allow DTA to search the space effectively [14].

3. DISCUSSION AND ANALYSIS

In these section we describe how to use DTA to optimize the statistic for SQL Server 2005 database, then we show the effect on that database. And it will ended by show the analysis of these process.

3.1 Optimization of SQL Server 2005 Query

For this experiment, we used the database and query loading from Microsoft SQL Server 2005 official training course number 2784A which titled Tuning and Optimizing Queries Using Microsoft SQL Server 2005.

The scenario is based on Baldwin Museum case study on chapter 3. Baldwin Museum of Science is a hands-on science center located in a medium-sized city in the eastern United States. Its mission is to educate the public and enrich local schools with knowledge of science and nature.

The museum recently upgraded its database to SQL Server 2005, and the database has now become an integral part of the day-to-day operation of the museum. Although the past two years have seen zgood museum growth, users have started to complain about the performance of this application.

The database named Baldwin2, will be tested with script. When the script is running without optimization from DTA, it consume about 3.35 min in execution time, we can see the detail of it in figure 4. And then we start the DTA, the user interface is in like figure 5. at the first DTA running, it will asked about the database and the load file. After that, we just click start analysis button like figure 6.

Then The DTA will show the result of its analysis based on SQL workload and database, like figure 7. Beside that, the DTA also give a recommendation based on SQL workload and database. These recommendation can be see in figure 8. in Figure 8 we can see that DTA is not only given recommendation for indexing the SQL Server database based on sql workload, but it can give recommendation for creating statistic too. There is 5 statistic that will be created for us if we apply these recommendation. Start from _dta_stat_197575742_7_8_5_6 until _dta_stat_2089058478_3.

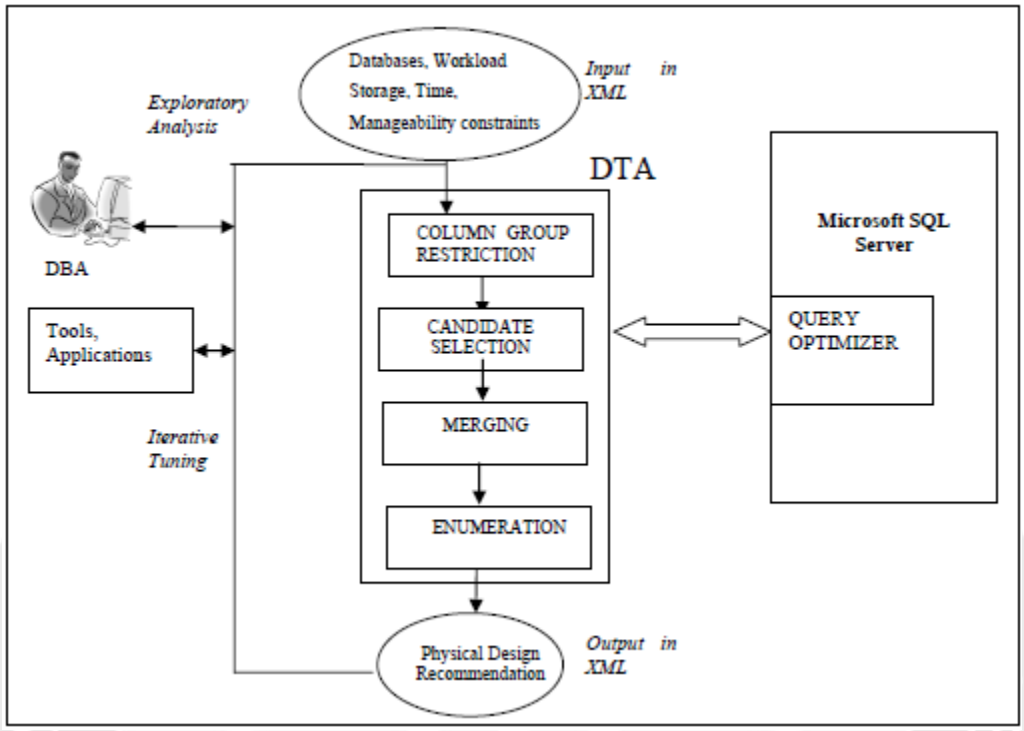


Figure 3. DTA Architecture

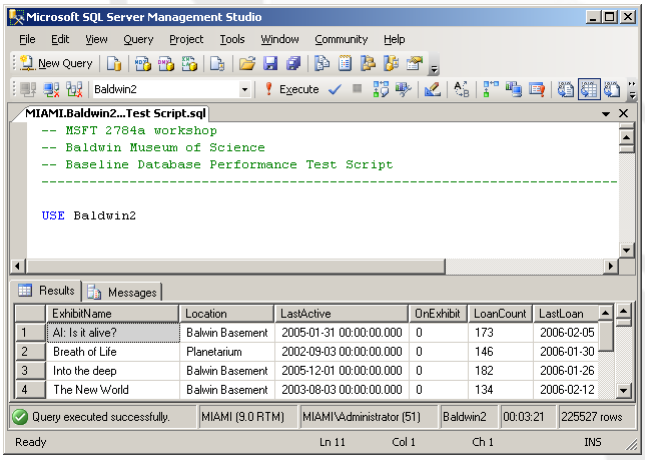


Figure 4. Load stress result without DTA

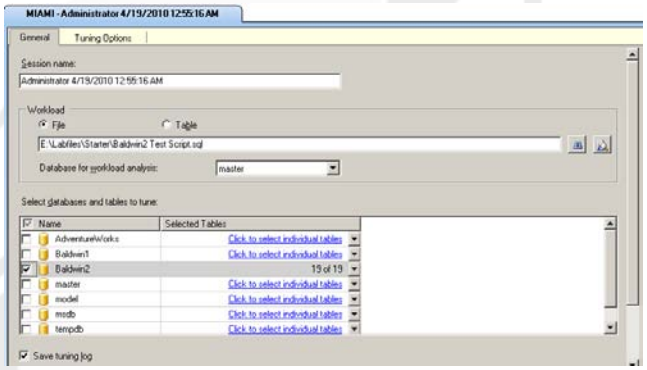


Figure 5. Load stress test file to test the database

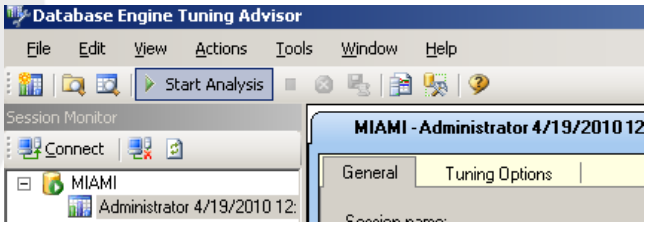


Figure 6. begin analysis

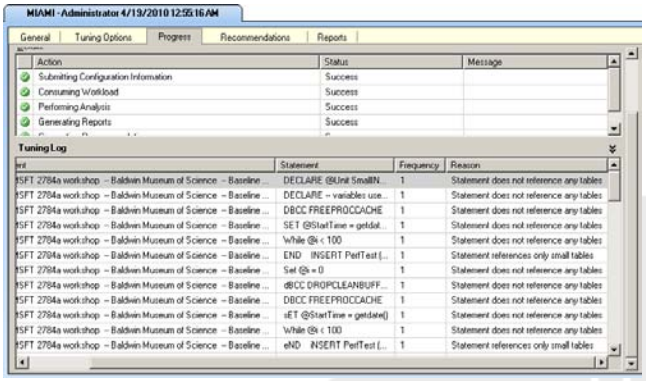


Figure 7. analysis result

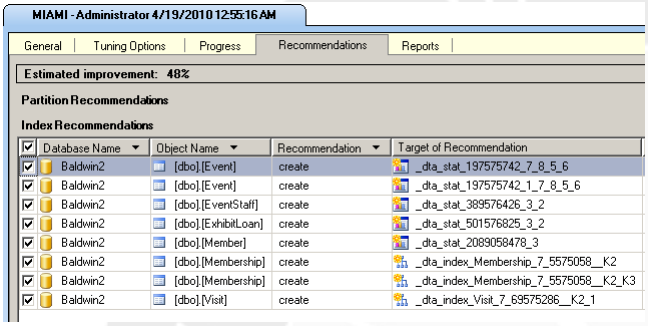


Figure 8. recommendation

After we applying the recommendation, figure 9 will showed, how many recommendation have been success in implementation.

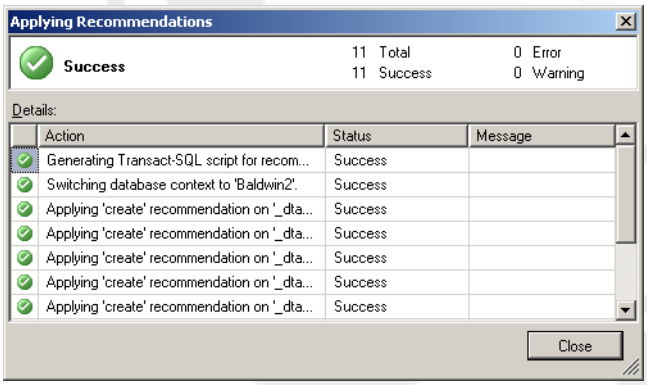


Figure 9. applying recommendation

If we success in applying recommendation, then the performance of baldwin 2 should be more better than before. Because of that, let try it by running the sql load script once more

It is shown in figure 10. it shown that the execution time is down to 1 minute 52 second, is far more efficient than the first time query run without optimization from DTA, which taken time 3 minutes 21 seconds. Based on these case study we can make a conclucion that using DTA the execution time can be sliced up to 50%

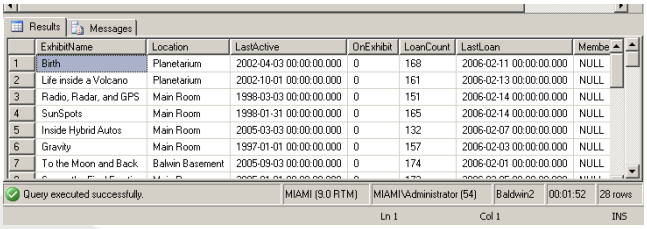


Figure 10. result from DTA

4. CONCLUSION

The purpose of this research is to prove that statistic has important role to optimize the SQL Server Database via Query Optimizer. Then we see that query optimizer is a part of DTA. So because DTA is a part of SQL Server 2005 Database tools , then we can make a conclusion that DTA is based on statistic too. It proven by DTA recommendation. It is no give recommendation just based on index, but it can give a recommendation based on statistic too..

5. REFERENCES

[1] Zhang.XS and Gosalia. A. 2008. Automatic Plan Choice Validation Using Performance Statistics. DB Test'08. (June , 2008)

[2] Chaudhuri. S, Agrawal.S, Kollar.L, Marathe.A, Narasayya.V, Syamala M. 2005. Database Tuning Advisor for Microsoft SQL Server 2005: Demo. ACM SIGMOD. (June. 2005), 930-932.

[3] L. Giakoumakis and C. Galindo-Legaria. Testing SQL Server's Query Optimizer: Challenges, Techniques and Experiences. *IEEE Bulletin of the Technical Comitee on Data Engineering*, 31, 1(March 2008), 37-44

[4] Chaudhuri. S, Bruno. N. 2002. Exploiting Statistics on Query Expression for Optimization. ACM SIGMOD. (June. 2002), 263-274.

[5] G. Graefe. The cascades framework for query optimization. *Data Engineering Bulletin*, 18(3), 1995.

[6] G. Graefe and D.J.DeWitt. The exodus optimizer generator. In *Proceedings of the 1987 ACM International Conference on Management of Data (SIGMOD'87)*, 1987.

[7] G. Graefe and W.J.McKenna. The volcano optimizer generator: Extensibility and efficient search. In *Proceedings of the Ninth International Conference on Data Engineering*, 1993.

[8] L.M. Haas, J. C. Freytag, G.M. Lohman, and H. Pirahesh. Extensible query processing in starburst. In *Proceedings of the 2000 ACM International Conference on Management of Data (SIGMOD'89)*, 1989.

[9] P.G. Selinger, M.M. Astrahan, D.D. Chamberlin, R.A. Lorie, and management system. In *Proceedings of the 1979 ACM International Conference on Management of Data (SIGMOD'79)*, 1979.

[10] S. Chauduri. An overview of query optimization in relational systems. In *Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*. ACM Press, 1998.

[11] Agrawal, S., Chaudhuri, S., Kollar L., Marathe A., Narasayya V. and Syamala, M. Database Tuning Advisor For Microsoft SQL Server 2005. In *Proceedings of the VLDB 2004 Conference*, Toronto, 2004.

[12] Agrawal, S., Chaudhuri, S. and Narasayya V. Automated Selection of Materialized Views and Indexes for SQL Databases. In *Proceedings of the 26th VLDB Conference*, Cairo, 2000.

[13] Chaudhuri, S. and Narasayya V. An Efficient, Cost-Driven Index Selection Tool for Microsoft SQL Server. In *Proceedings of the 23rd VLDB Conference*, Athens, 1997.

[14] Agrawal, S., Narasayya V. and Yang, B. Integrating Vertical and Horizontal Partitioning Into Automated Physical Database. In *Proceedings of the ACM SIGMOD*, Paris, 2004

Spatial Autocorrelation Modelling for Determining High Risk Dengue Fever Transmission Area in Salatiga, Central Java, Indonesia

Sri Yulianto J.P.
Faculty of Information Technology
Satya Wacana Christian University
Jl. Diponegoro 52-60, Salatiga
50711, Indonesia
sriyulianto@gmail.com

Kristoko Dwi Hartomo
Faculty of Information Technology
Satya Wacana Christian University
Jl. Diponegoro 52-60, Salatiga
50711, Indonesia
kristoko@gmail.com

Krismiyati
Faculty of Information Technology
Satya Wacana Christian University
Jl. Diponegoro 52-60, Salatiga
50711, Indonesia
blesschris@gmail.com

ABSTRACT

This study develops Autocorrelation Spatial Model as a software for mapping high risk Dengue fever transmission area. The data used in this study are the cases of dengue fever reported to the Health Department in Salatiga from 1998 to 2008. It also uses the number data of mosquito larva free of Salatiga. The data analysis method used in this study s Global Autocorrelation represented by Moran's index and Local Indicator Spatial Association (LISA) represented by cluster map and significance map. The result of the study shows that the Moran Index is -0.0251 and -0.0049 which describes the association between variables which is negative towards population density variable and mosquito larva free number (abj) towards dengue fever case. *Hotspots* (positive spatial association area) in Salatiga and Gendongan have an impact towards the hotspot formation in Mangunsari, Sidorejo Lor and Kutowinangun in 2006, Kelurahan Sidorejo Lor and Tegalrejo in 2007; Mangunsari and Gendongan in 2008. The result of the study using LISA has not described that dengue fever case is always significant towards the increase of abj variable value. Based on the map of high risk dengue fever transmission modelled with spatial autocorrelation, then preventive and anticipative steps could be taken due to the possibility of the spreading area having the high risk potential of dengue fever transmission.

Keywords

spatial autocorrelation, dengue fever, spatial modelling

1. BACKGROUND

Mathematics modelling has been widely used for studying the dynamic of a system having variables with high complexity in some areas such as chemistry, physic, pharmacy, and medicine. Modelling works using concepts of material, individual and energy quantity changes in the variable constructing the system. The modelling will be dynamic when it uses time as its independent variable in differential equation. [1]. Modelling of phenomenon and detection of spatial object clustering is an important function spatial statistic. The objectives are; firstly it is for detecting and quantifying the location of object clustering as the source of disease transmission, the source of the disease or other phenomena. Secondly, it is for knowing the relation between particular object cluster in an area and its surrounding. [2][3]. The mathematic modelling that has been applied is using spatial stochastic concept for simulating the spreading pattern of animal disease in North America[4]. Another implementation is spatial and temporal pattern modelling for transmission of infection sources of schistochomyacea which is endemic in XiChang, China[5]. Apart from that, ther is also Bayesian Modelling for mapping the mortality data heterogeneity caused by cancer in Germany [6]. This study focuses on the implementation of spatial auto correlation as an indicator for modelling the area of high risk dengue fever transmission in Salatiga, Central Java, Indonesia. Basically, spatial autocorrelation works by comparing activity of an object in an area having characteristic resemblance to other activity or object in other areas and their surrounding. The comparison of those two sets of attribute with similar resemblance will result in a spatial pattern called as positive spatial correlation [7]. The activity or object being assessed important for determining the high risk dengue fever transmission area are climatology characteristic, mosquito larva free number, knowledge attitude survey society behaviour and population density. Those variable values significantly have relation to their surrounding. The relation pattern formed could be used as an indicator for determining the degree of the risk of disease transmission in that area. Based on the map of high risk dengue fever transmission modeled using spatial autocorrelation, preventive and anticipative steps could be organized for the possibility of the spreading of area having high risk potential of dengue fever transmission [8].

2. RESEARCH METHOD

The data used in this study is a case study of dengue fever reported to the health department of Salatiga from 1998 to 2008. Furthermore, it uses rainfall data from 1970 to 2009. It also uses population density data of Salatiga from 2001 to 2008. The mosquito larva-free number data of all area in Salatiga is also used [8]. The case of dengue fever takes places in Saltiga from 1998 to 2008 could be summarized in picture 1.

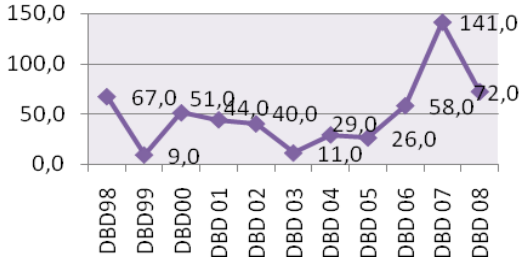


Figure 1. Dengue Fever case from 1998 t0 2008 in Salatiga

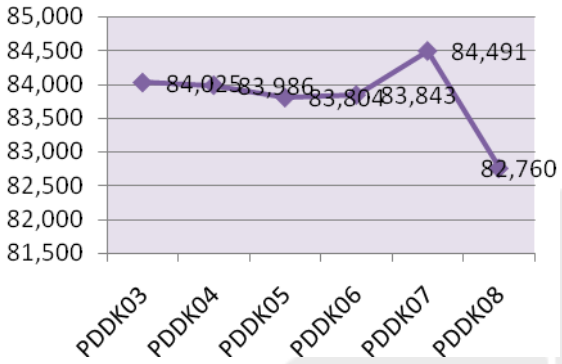


Figure 2. The population of Salatiga in 2003 to 2008.

The data of musquito larva-free number from 1998 to 2008 could be seen in Figure 3.

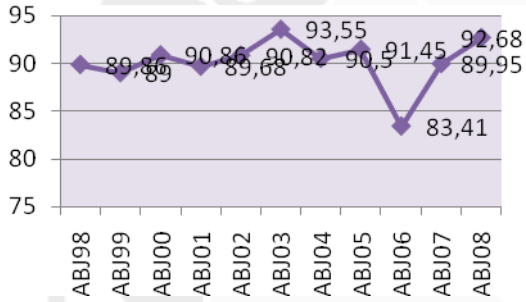


Figure 3. Data of musquito larva-free number from 1998 to 2008

The data analysis technique is done using spatial association comprising of *Global Spatial Association* for examining the pattern formed frm the variables of object in a particular area and its surrounding. Meanwhile, *Local Spatial Association* is used for examining the pattern formed and information obtained from a variable or object in a particular area. Nevertheless, the pattern formed between *Local Spatial* and *Global Spatial* does not have any linear trend [9]. *Global Spatial Autocorrelation* represented by Moran's Index is as follows :

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} z_i z_j}{\sum_{i=1}^{n} z_i^2} \quad (1)$$

In the above equation, value is the the value of variable set of x in the study area of n. wij is the relation between data in study rea of I and j[10].

Local Spatial Association is used for identifying of spatial pattern in a study area. *Local Moran's* could be represented as follows :

$$I = \frac{zi \sum_{j=1}^{n} wij zj}{zi \sum_{j=1}^{n} z_j^2 / n} \quad (2)$$

For i=1,...n. The value of Ii is positive shown by the local cluster formation in i surrounding area. The function of *Local Moran's* could be used for showing the data instability like local data deviation from the spatial assocation global pattern or hotspots identification [10].

3. RESULT OF THE STUDY AND DISCUSSION

Global Spatial Autocorrelation is analyzed using Moran'I and the result is visualized in Moran Scatter plot. Moran Scatter plot describes the association between the population density variable towards the number of dengue fever cases in 2003 – 2008 in Salatiga as shown in Picture 5. Based on the data in Picture 1 and 2, the biggest number of the those suffer from the diseases and the highest number of the population is in 2007. Global spatial autocorrelation in 2007 shown in Figure 4.

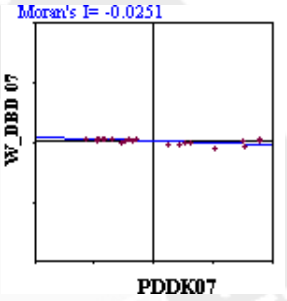


Figure 4. *Global spatial autocorrelation* of population density in 2007

Data in the 1st and 2nd quadran indicates that there is a positive spatial association in the study area. It tells that the density population significantly causes the case of dengue fever in the study area. The study area is positive spatial which means that there is high characteristic similarity of variable value causing the dengue fever case in the study area and its surrounding. It could also mean that the value of the cause of dengue fever case variable in the study area is low but surrounded by study area with high value. Positive spatial association indicates that there is clustering of dengue fever case caused by same study area variable; population density. The Moran's Index of -0.0251 describes the negative association between variables. Therefore factor of population density towards the dengue fever case in 2007 does not have any significant correlation seen from global autocorrelation.

In a small number, there are data in 3rd and 4th quadrants. This indicates that there is negative spatial association. In the study area there is characteristic difference so that the population density as the cause of dengue fever is not significant. It explains that the variable value characteristic similarity of the value of population density variable as the cause of dengue fever case is low. In other words, it means that the similarity of the value characteristic of population density variable as the cause of the dengue fever case in the study area is high and surrounded by an area of study with low similarity of characteristic value. The negative spatial association indicates that there is clustering of dengue fever case caused by local characteristic with different variables. When compared to the mosquito larva-free number, it could be represented with Moran's index in Scaterplot as seen in picture 5. Data are distributed evenly in those four quadrants. Part of the mosquito larva free number is negative spatial association and another part is positive. Moran's index value is -

0.0049 describing the association between variables is negative. Therefore, mosquito larva free number and dengue fever case in 2007 do not have any significant relationship seen from global autocorrelation.

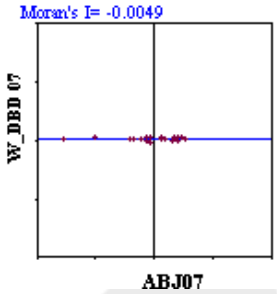


Figure 5. *Global spatial autocorrelation* of mosquito larva free number in 2007

The analysis of *local indicator spatial association* (LISA) basically has two criteria. Firstly, each data indicates the *significant spatial clustering* in the study area. Secondly, all data are proportional as global spatial association indicator. LISA is implemented to detect the clustering around the data which is significant spatial.

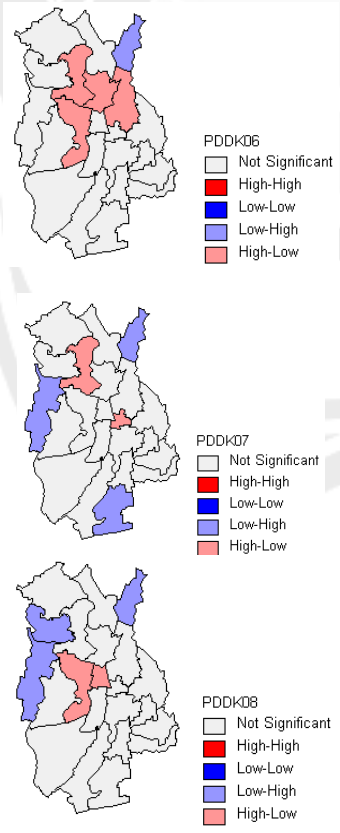


Figure 6. LISA analysis on population density variable towards the number of dengue fever cases.

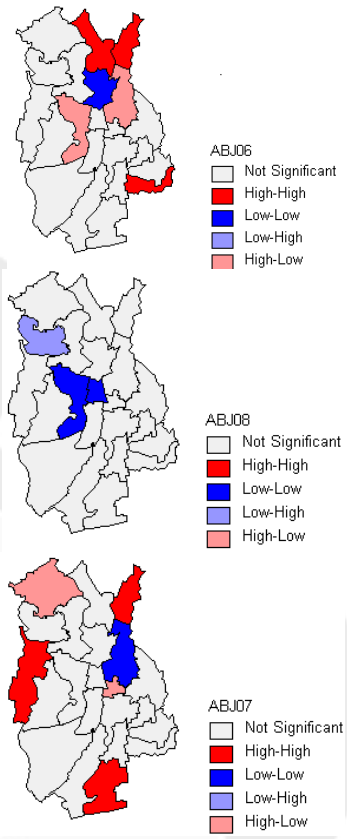


Figure 7. LISA analysis between mosquito larva free number and the number of the cases.

Dengue fever cases is less significantly affected by mosquito larva free umber. In 2007, the mean of mosquito larva free number is 89,95%. It is far away from the secure value of > 95%[11]. Nevertherless, it has the highest number of dengue fever cases in the last 10 years which is 141 cases. Meanwhile, when compared to 2006, the mean of mosquito larva free number is 83.41% , and it is the lowest in the last 10 years. There were 83 dengue fever cases. (Figure 7).

4. CONCLUSION

Map of high risk dengue fever transmission could be developed using spatial autocorrelation model as preventive and anticipative steps towards the possibility of the spreading of area having high risk dengue fever transmission potential. The value of Moran's index of -0.0251 describes the association between variable which is negative. Therefore, population density factor and dengue fever case in 2007 does not have any significant correlation seen from the global autocorrelation. Moran's index of -0.0049 describes the association betweeb variables is negative so that mosquito larva free number factor and the dengue fever cases in 2007 does not have significant correlation seen from the global autocorrelation. L*ocal indicator spatial association* (LISA) analysis describes that in 2006-2008, *hotspots* in Salatiga and Gendongan have impact towards *hotspots* formation in Mangunsari, Mangunsari, Sidorejo Lor and Kutowinangun in 2006; Sidorejo Lor and Tegalrejo in 2007; and Mangunsari and Gendongan in 2008. Precisely, it has not been proved that dengue fever cases is always significant

towards the increase of variable value of mosquito larva free number. In 2007 the mean of mosquito larva free number is 89,95%, but it has the biggest number of dengue fever cases in the last 10 years; it is 141 cases. When compared to 2006, the mean of mosquito larva free number is 83.41% , it is the lowest value in the last 10 years having 83 cases.

5. REFERENCES

[1] Soetaert Karline dan Thomas Petzoldt, 2009, *Inverse Modelling, Sensitivity and Monte Carlo Analysis in R Using Package*, FME Netherlands Institute of Ecology, Technische Universitat at Dresden.

[2] Zhang Tonlin dan Ge Lin, 2008, *Spatial scan statistics in loglinear models*, Computational Statistics and Data Analysis 53 (2009) 2851-2858, Department of Statistics, Purdue University, West Lafayette, IN, USA.

[3] Anselin Luc, Ibnu Syabri dan Youngihn Kho, 2005, *GeoDa: An Introduction to Spatial Data Analysis*, Geographical Analysis 38 (2006) 5–22, Department of Geography, University of Illinois, Urbana-Champaign, Urbana.

[4] Harvey Neil, Aaron Reeves, Mark A. Schoenbaumc, Francisco J. Zagmutt-Vergara, Caroline Dube, Ashley E. Hill, Barbara A. Corso e,W. Bruce McNab, Claudia I. Cartwright, dan Mo D. Salman, 2007, *The North American Animal Disease Spread Model: A simulation model to assist decision making in evaluating animal disease incursions*, Preventive Veterinary Medicine 82 (2007) 176–197,USA.

[5] Xu Bing dan Peng Gong, 2006, *Spatial Temporal Modeling of Endemic Diseases: Schistosomiasis Transmission and Control as an Example* , Center for Natural and Technological Hazards, Department of Geography, University of Utah, Salt Lake City, UT 84112, USA

[6] Knorr-Held Leonard dan Nikolaus Becker, 1999, *Bayesian Modelling of Spatial Heterogenity in Disease Maps with Application to German Cancer Mortality Data*, Institut fur Statistik, Universitat Munchen, Germany

[7] Ceccato Vania dan Anders Karlstrom, 2001, *A new information theoretical measure of global and local spatial association*, Dept of Economics, University of California Berkeley, Evans Hall, CA–94720 Berkeley, USA

[8] Yulianto S, Kasmiyati S,Marina M,Hartomo KD, 2008, *Pengurangan Potensi bencana epidemi, wabah dan KLB beberapa penyakit tropis melalui penerapan paradigma pengurangan resiko yang diintegrasikan dengan kurikulum pembelajaran pada sistem manajemen bencana*, Laporan Hibah Kompetitif Batch IV, Direktorat Pendidikan Tinggi, Departemen Pendidikan Nasional.

[9] Karlstrom, Anders, Ceccato dan Vania, 2000, *A new information theoretical measure of global and local spatial association*, Royal Institute of Technology, Sweden

[10] Scrucca Luca, 2005, *Clustering multivariate spatial data based on local measures of spatial autocorrelation*, Dipartimento di Economia, Finanza e Statistica Universit`a degli Studi di Perugia, Italy

[11] Santoso dan Anif Budiyanto, 2009, *Hubungan Pengetahuan Sikap dan Perilaku Masyarakat Terhadap Vektor DBD di Kota Palembang Prov. Sumatera Selatan*, Jurnal Ekologi Kesehatan Vol. 7 No. 2, Agustus 2008 : 732 – 739.

Supply Chain Improvement with Design Structure Matrix Method and Clustering Analysis (A Case Study)

Tanti Octavia
Department of Industrial Engineering Faculty of Industrial Technology Petra Christian University Siwalankerto 121-131 Surabaya 60236, Indonesia
+62-31-2983433
tanti@petra.ac.id

Siana Halim
Department of Industrial Engineering Faculty of Industrial Technology Petra Christian University Siwalankerto 121-131 Surabaya 60236, Indonesia
+62-31-2983433
halim@petra.ac.id

Stefanus Anugraha Lukmanto
Department of Industrial Engineering Faculty of Industrial Technology Petra Christian University Siwalankerto 121-131 Surabaya 60236, Indonesia
+62-31-2983433

Harvey Sutopo
Department of Industrial Engineering Faculty of Industrial Technology Petra Christian University Siwalankerto 121-131 Surabaya 60236, Indonesia
+62-31-2983433

ABSTRACT

Many strategic businesses attempt to achieve coordinating operations of company across departments using information and communication flow for their supply chain network. One of customer goods companies in Surabaya attempts to improved their flow of information and communication using Design Structure Matrix (DSM). DSM is a method that could provide an alternative system grouping work activities. The activities of each department data and data interaction between the elements are needed to develop DSM. The research is done in the inter-department and intra department. The analysis technique used is the clustering analysis, which consists of hierarchical methods. The results show that for inter-department single linkage hierarchical clustering method with a number of groups of three is the best number of groups. For intra-planning department and intra-RMS department, the best number of groups is 14 and 8, respectively, using ward linkage method.

Keywords

supply chain, the design structure matrix, clustering analysis.

1. INTRODUCTION

Nowadays, competitive pressures and changes in the economic conditions have forced companies to continuously improve their competitive advantage by creating new strategic business. Many strategic businesses attempt to achieve coordinating operations of company across departments using information and communication flow for their supply chain network. Simchi (2005) stated that coordination of the supply chain has become strategically important as new forms of organization, such as virtual enterprises, global manufacturing and logistics networks, and other company-to-company alliances, evolve.

The customer goods industries are not an exception for developing and creating the new strategies. They usually have a long supply chain and a complex network of supply chain. The multifaceted of supply chain network may occur the ineffective of information flow, inefficient the use of information and communication. Ogulin (2003) suggests three distinctive waves of supply chain management in the new economy: operational excellence, supply chain integration and collaboration, and virtual supply chains. Operation excellence refers to the degree of sharing within company, workflow activities across department within the company in order to achieve efficiencies from increased order accuracy and timely shipments. Workflow activities and the interactions between elements can be depicted in a design structure matrix (DSM). A DSM can achieve an alternative system to perceive how strong the relationship between the elements effectively. After developing a DSM, the closeness relationship of activities in a DSM could be clustered using the use of information and communication. The clustering analysis is useful to classify the groups with the similar characteristics (Barolomei, 2007).

This research aims to propose an alternative system in a customer goods industry by applying DSM and clustering analysis.

2. LITERATURE REVIEW

Design structure matrix is a matrix which aims to show all the interactions between elements (Chen & Huang, 2007). DSM has the advantage that they can improve the structure of the system by using matrix-based analysis techniques. Figure 1 presents the structure of the DSM. The input on a cell is the relationships between two elements.

	a	b	c	d	e	f	g	h
a			x					x
b					x	x	x	
c								
d							x	x
e			x				x	
f		x						
g						x		
h				x	x			

Figure 1. The example of DSM

Type of interaction in DSM can be divided into two, namely numerica l binary DSM (Chen & Huang, 2007). The first type is the type of interactions that binary interactions, which interaction is only worth or not there is interaction. This type of interactions is able to show interaction between each element, but still have shortcomings. This type cannot describe how strong the interaction between one and another element. The second type of interaction in the DSM is the numerical which the value is worth its interaction with the figures.

2.1 Clustering Analysis

Clustering analysis is a method of classifying an object into one or more than one group, so that each object is located in one group will have the same value of interaction. Clustering analysis aims to form groups with similar characteristics. Two kinds of methods in clustering analysis are hierarchical methods and non-hierarchical method (Sharma, 2006). Hierarchical method is a method that takes into account the distance between the two groups. Five-way hierarchical clustering methods are in the following.

- Single linkage clustering
- Complete linkage clustering
- Centroid linkage clustering
- Average linkage clustering
- Ward linkage clustering

In order to calculate the similarity value, the squared Euclidean distance can be applied. The squared Euclidean can be calculated in the formula 1.

$$Dij = \sum_{k=1}^{p} (X_{ik} - X_{jk})^2 \quad (1)$$

where:

Dij: distance between elements i and j

X: the different data elements on

i: an element which was in line

j: is the element in column

k: a number of variables of each of the elements

3. RESEARCH METHODOLOGY

This research was designed and conducted using primary and secondary data. Primary data is applied by doing interview to the manager and his subordinates in the planning department and raw material store. The interviews used to obtain the workflow for each department. In addition, it also gives the information for determining the elements that are based on the activity manager and the subordinates. Secondary data used there are two that work instructions and past data on program systems and applications products in data processing in order to add elements that are not derived from the interviews and obtained the data flow.

Data collection is designed in a Design Structure Matrix (DSM) and clustered using hierarchical methods and non-hierarchical method. Hierarchy has five different methods of linkage which often used for complete linkage, single linkage, average linkage, wards, and centroid. These five methods will be selected based on the highest similarity value.

Clustering analysis aims to classify the activities contained in the DSM with a number of specific groups. Grouping is done based on distance data, which will make the flow of information between departments optimally. The method used to determine which group has a high value and the closeness low in the analysis of the distance is squared Euclidean distance. Finally, selection the best method of clustering analysis is done by considering the current conditions.

4. ANALYSIS

After collecting data, design structure matrix (DSM) is built. The activities of two departments can be classified into 129 elements. The interaction values in each cell are obtained from number of transactions in raw material store department and number of daily activities in planning department. The example of DSM is shown in figure 2.

Element	T-1	T-2	T-3	T-4	WOP-1	WOP-2
T-1	-	756	0	0	0	0
T-2	4190	-	112	0	0	0
T-3	0	112	-	0	0	0
T-4	3895	756	108	-	105	0
WOP-1	0	0	0	0	-	0
WOP-2	0	0	0	0	0	-

Figure 2. The example of DSM

Improving supply chain in this research is done by classifying the activities that have the same number of interactions (in one group). A group is expected to enlarge the company performance since they can communicate and inform the information effectively. Clustering analysis is accomplished using Minitab software. Clustering methods used there are two, namely, hierarchical methods and non-hierarchical method. Hierarchy has five different methods of linkage which often used for complete linkage, single

linkage, average linkage, wards, and centroid. These five methods will be selected based on the highest similarity value. The similarity level of each method and each clustering can be seen in table 1.

The result shows that single linkage method gives the highest similarity value. After calculating single linkage, the best clustering is determined through a combination of the computation RMSSTD with the company's current condition.

Similarity Hierarchical method is an appropriate method because all elements have relationships with one another. Value of RMSSTD for each number of groups can be seen in table 2. The result shows that 10 groups give the smallest RMSSTD. But, this classification does not fit with the company's condition and consideration. After discussing and interviewing the company's expert about the classification each number of groups, we can get the result that the best number of cluster is three groups.

In this research, clustering analysis is also applied to group the activities in Planning Department and Raw Material Store Department. The number of the selected group in Planning Department is 14. The result shows the single linkage clustering method is the highest in term of similarity value. Unfortunately, it is not suitable for company's current condition. Wards linkage clustering method gives an appropriate number of groups in term of company's current condition (figure 3). Group activity was initially assessed based on the type of product, whereas proposed group is classified based on the closeness activities of the group.

Table 1. Similarity Level using Hierarchical Methods

Cluste Ring	Single linkage	Centro id	Complete	Averag e	Ward
1	80.673	58.28	0	40.067	-285.258
2	86.798	76.163	52.809	73.596	27.457
3	86.798	79.829	58.985	78.388	35.904
4	86.798	83.247	60.394	79.178	51.162
5	86.798	83.497	62.63	82.346	60.114
6	86.798	86.798	73.538	86.546	60.394
7	90.827	86.798	81.513	86.798	77.566
8	90.899	86.852	86.798	86.798	82.105
9	92.919	88.221	86.798	86.798	86.646
10	93.166	90.911	86.798	90.337	86.798

Table 2 Value of RMSSTD from 1 to 10 groups

Number of groups	RMSSTD
1	451.1007
2	440.1591
3	394.4475
4	350.8927
5	310.8812
6	276.2852
7	250.2835
8	232.0255
9	216.6936
10	208.5315

For Raw Material Store Department, the single linkage clustering method gives the highest similarity value. Indeed, it is not appropriate with the company's condition. Therefore, wards linkage clustering with number of groups is eight can be applied in term of company's condition. Currently, the group activity is classified based on the early function of each part (receiving, storing, shipping, etc.). The proposed group attempts to combine the administration activities on any part of the RMS.

5. CONCLUSION

The result of clustering analysis interdepartmental planning and RMS differs from the grouping prior to the DSM. Total group originally owned by the company prior to using the DSM is the eight groups, after performing clustering analysis with the DSM has been reduced to three groups. The closeness relationship between the planning department and department RMS makes both departments need to be placed together or into one large department.

Clustering analysis of intra-departmental planning has brought changes in the group activities held by the department. The first group owned by the department is planning three groups, after performing clustering analysis of these groups has increased to fourteen groups.

Analysis of intra-departmental grouping of RMS has brought changes in the group activities held by the department. Total group originally owned by the department RMS are five groups, after analyzing the grouping of these groups has been increased to eight groups.

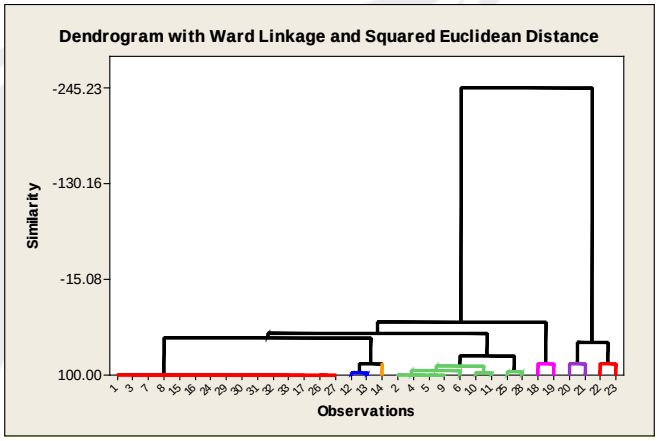


Figure 3. The Example of Dendogram with Ward Linkage method and Squared Euclidean Distance.

6. REFERENCES

[1] Bartolomei, Jason E. (2007, June). Qualitative Knowledge Construction for engineering systems: Extending the design structure matrix methodology in scope and procedure. "Massachusetts Institute of Technology, 1-191

[2] Chen, Gary. & Huang, Enzhen. (2007, April). A systematical approach for supply chain design using matrix structure. Montana State University, 285-299

[3] Ogulin R. Emerging requirements for networked supply chains (2003). In: Gattorna JL, Ogulin R, Reynolds MW, editors. Gower handbook of supply chain management. Burlington, VT: Gower Publishing; 2003. p. 486–500.

[4] Simchi-Levi, D., & Kelompoknsky, P. (2008). Designing and managing the supply chain: concepts, strategies and case studies. New York: McGraw-Hill.

[5] Sharma, Subhash. (2006). Applied multivariate techniques. Canada: John Wiley and Sons, Inc.

The Comparation of Similarity Detection Method on Indonesian Language Document

Anna Kurniawati
Gunadarma University
Jl. Margonda Raya No 100
Depok, Indonesia
021-78881112
ana@staff.gunadarma.ac.id

Lily Wulandari
Gunadarma University
Jl. Margonda Raya No 100
Depok, Indonesia
021-78881112
lily@staff.gunadarma.ac.id

I Wayan Simri Wicaksana
Gunadarma University
Jl. Margonda Raya No 100
Depok, Indonesia
021-78881112
iwayan@staff.gunadarma.ac.id

ABSTRACT
Six semester student at the Gunadarma University obliged to make the writing of scientific or scholarly research, enabling students to take writing materials owned by others. To overcome these problems, it is not enough simply to remind the students that such action is not good. Detection of text document similarity is one solution that should be done so that fraudulent activity can be minimized.

Systems or tools to detect plagiarism is quite a lot, both to detect the text or document or to detect the source code programming. Systems or tools to detect the similarity of documents in English have been widely developed. Examples of tools that have been developed for English-language document are Turnitin, Eve2, CopyCathGold, WoedCheck, Glatt, Moss and Jplang and so forth. Similarity detection research done on text documents or documents written in Indonesian text was still relatively little done.

In this paper will present an analysis of several methods to detect similarities documents written in Indonesian. The methods used are the keyword method, the Karp and Rabin method and Jaro Winkler Distance method.

Data used for this test is abstraction of scientific writing data Gunadarma University Information Systems majors. The data abstraction will be modified into three kinds of abstraction: first, there was only a partial abstraction of the same data abstraction will be compared. Second is the abstractions that sentence changed positions from data abstraction are compared and the third is abstraction whose content is replaced with a synonym of the word in the abstract that will be compared. From the test results, the best method is the method of Rabin Karp, except for a synonymous.

Keywords
Indonesian language, Detection, Document, Similarity

1. INTRODUCTION
Academic communities specially student is very enabled conduct writing or research that take materials of others property writing, because the development of information technology. The development of information technology that provides the facility to copy and modify the text (copy and paste) and facilities that allow connection to access other people's work for free through the Internet, can facilitate to take other people's without mention owners of original data source.

Six semester student at the Gunadarma University obliged to make the writing of scientific or scholarly research, enabling students to take writing materials owned by others. To overcome these problems, it is not enough simply to remind the students that such action is not good. Detection of text document similarity is one solution that should be done so that fraudulent activity can be minimized.
Systems or tools to detect plagiarism is quite a lot, both to detect the text or document or to detect the source code programming. Systems or tools to detect the similarity of documents in English have been widely developed. Examples of tools that have been developed for English-language document are Turnitin, Eve2, CopyCathGold, WoedCheck, Glatt, Moss and Jplang and so forth. Similarity detection research done on text documents or documents written in Indonesian text was still relatively little done.
In this paper will present an analysis of several methods to detect similarities documents written in Indonesian. The methods used are the keyword method, the Karp and Rabin method and Jaro Winkler Distance method.
This paper is divided into four parts, namely the one describing the introduction, section two describes plagiarism, including definitions of plagiarism, types of plagiarism, an engineering approach to plagiarism detection, plagiarism detection methods and existing tools for plagiarism detection. In part three will be presented on the test was conducted on the sample test data, test steps and test results. In the four presented the results of testing and analysis, while the fifth section exposed on the conclusions of the experiments which were conducted in three parts.

2. SURVEY IN DOCUMENT SIMILARITY MEASUREMENT

2.1. Plagiarism

2.1.1. Definition of Plagiarism
Plagiarism is removal of essays, opinions, etc. from others people and make it as their own articles and opinions [6]. Plagiarism can be considered a criminal offense for stealing the copyrights of others. In the world of education, principals of plagiarism can have severe penalties such as expelled from school / university. Someone who conducts Plagiarism is called plagiarist.
Plagiarism can be Classed as follows: [4]

a. Using the writings of others without giving a clear sign (for example, by using quotation marks or block different paragraphs) that the text is taken directly from the writings of others.

b. Taking the ideas of others without giving sufficient annotation of the source.
c. Acknowledging the findings of others as one's own.
d. Acknowledging the work of the group as a possessive or a result of his own.
e. Presenting the same papers in different occasions without mentioning its origin.
f. Summarize and make paraphrase (indirect quote) without mentioning its source.
g. Summarize and make paraphrase with reference to the source, but a series of sentences and the choice of word was too similar to its source.

The things that are not classified as plagiarism are as follows: [4]

a. Using the information in the form of general facts.
b. Writing back (by changing a sentence or paraphrasing) the opinions of others by giving a clear source.
c. Quoting the writings of others to taste with clearly marking out the passage and write down the source.

2.1.2. Plagiarism in the academic area

Besides problem usual plagiarism, swaplagiarisme also often happens in the world of academic. Swaplagiarisme is to reuse part or all of the author's own work without giving the original source. Plagiarism in the academic field can be divided into two, namely: [10]

Content-based file comparison

Content-based file comparison approach is appropriate approach to the text as a student essay assignment.

Content-based comparison of source code

This approach is used to detect plagiarism for the source code programming.

2.2. Plagiarism Detection

In this section will be explained about plagiarism detection approaches and research studies have been done to the plagiarism detection approaches. Plagiarism detection approaches can be categorized into three categories such as substring matching, keyword similarity and fingerprint. Here's an explanation of each of these approaches. [3]

2.2.1. Substring Matching

Substring matching approach is an approach to identify the same string that is used as an indicator for plagiarism. In this approach, substring is described in suffix trees and graph that is used to take part of plagiarism. One of the algorithms used in the substring matching approach is the Jaro-Winkler algorithm. Here is an explanation of the algorithm.

Jaro-Winkler Algorithm [12]

Jaro-Winkler distance is a variant of the Jaro distance metric is an algorithm to measure similarity between two strings, this algorithm is usually used in duplicate detection. The higher the Jaro-Winkler distance for two strings, the more similar to that string. Jaro-Winkler distance is the best and suitable for use in the comparison of short strings such as names of people. The normal score of 0 indicates no similarity, and one is exactly the same. Jaro-Winkler algorithm time complexity quadratic distance has a runtime complexity that is very effective in the short string and can work faster than the edit distance algorithm.

The basis of this algorithm has three parts are:

1. Calculate string lengths,
2. Find the same number of characters in two strings, and
3. Find the number of transpositions.

2.2.2. Keyword Similarity

The principle of this approach is to extract keywords from the document and then compared with the keyword in the document stated. If the similarity exceeds the threshold, the document will be divided into smaller parts, which will then be compared recursively. This approach assumes that the plagiarism usually occurs in a similar document.

2.2.3. Fingerprint Analysis

The most popular approach to analyzing text plagiarism is detected sequences that overlap with the way fingerprint. The document is divided into sequences, called chunks, from the reading of digital documents is calculated documents pattern. When reading a document pattern, it is inserted into the hash table. Banging show an appropriate sequence. One algorithm used in fingerprint analysis approach is the Karp-Rabin algorithm. Here is an explanation of the algorithm.

Karp-Rabin Algorithm

Karp-Rabin algorithm uses a hash function that provides a simple method to avoid the time complexity O (m2) in many cases. Instead of checking the position of each pattern contained in the text, would be more efficient if done checking only on the desired pattern. Checking the similarity between two words using a hash function.

To further assist in string matching problem, the hash function shall have the following properties [11]:

1.Capability of efficient computing.

2.Diskriminasi high against the string.

3. The function hash (y [j +1 .. j + m]) must be easily dikomputasi from

- Hash (y [j .. j + m-1])
- Hash (y [j + m])

Karp-Rabin algorithm has the following characteristics [11]:

• Using a hash function
• Preprocess phase in the time complexity O (m) and place a constant.
• Phase searches in time complexity O (mn)
• O (n + m) estimates the current time

3. METHOD

To determine the most appropriate method to detect similarity of document in Indonesian, then conducted testing to methods that are already exist. The first method to be tested is the keyword method, the second method is a method of fingerprinting document with Karp Rabin algorithm, the third method is a string matching method with the Jaro Winkler Distance algorithm and fourth method is manual method.

Data that used for this testing is scientific writing abstraction data as many as three abstraction. Each abstraction data will be modified become three kinds of abstraction that is:

1. Abstraction that its contents just part of in common from abstraction data that will be compared.

2. Abstraction that its sentence position altered from abstraction data that will be compared.
3. Abstraction that its contents changed with synonym from word at abstraction that will be compared.

Scenario of testing that will be done is the three of abstraction document that has been modified, will be looked for percentage of document similarity in comparison with document of original abstraction. Searching of document similarity will be done by using four methods that are keyword methods, Karp Rabin method, Jaro Winkler Distance method and manual method.

4. TESTING

4.1 Testing Data

Testing Data has been explained at part previously. The following are data example that used in testing, that are:

1. Abstraction is only part of the same contents of abstraction data to be compared.
2. Abstraction is the position of the sentence was changed from abstraction data to be compared.
3. Abstraction whose content is replaced with a synonym of the word in the abstract that will be compared.

Examples of data used will be presented in the figure below.

1. Original Abstraction Document.

This Original abstraction document is a document that will be compared with other abstraction document. In this document there are seven sentences of abstraction, as shown in Figure 1 below.

Judul

APLIKASI PENJUALAN BARANG PADA TOKO ITSARO
DENGAN NETBEANS 5.5 DAN MYSQL

Abstraksi

Proses transaksi dalam kegiatan penjualan yang cepat dan tepat merupakan faktor yang sangat penting dalam pengoperasian suatu perusahaan. Sehingga diperlukan suatu aplikasi penjualan yang mampu mengolah transaksi secara efektif dan efisien agar dapat meningkatkan pelayanan terhadap pelanggan atau konsumen dan peningkatan pendapatan perusahaan. Dalam penulisan ini penulis membuat aplikasi penjualan menggunakan Netbeans 5.5 dan MYSQL server 5.0 untuk pengolahan basis datanya. Dalam pembuatan laporan, iReport 2.0.2 mendukung penulis untuk menentukan jenis laporan yang diinginkan. Penulis mendapat data yang dibutuhkan dengan mengunjungi Toko Itsaro yang terletak di Tambun, Jawa Barat. Studi pustaka yang dilakukan penulis adalah dengan mencari dan mengutip teori – teori dari buku-buku dan internet. Setelah melakukan analisa terhadap aplikasi penjualan ini dapat disimpulkan bahwa aplikasi ini selain membantu kelancaran usaha perusahaan juga dapat membantu manajemen perusahaan dalam pengambilan keputusan yang tepat.

Figure 1. Original Abstraction Data

2. Abstraction Document Modification 1

This Abstraction document 1 is a modification document of the original abstract document. Modifications are done is take some sentences contained in the original document and added a sentence of abstraction different from the original document abstract. Sample document modification of abstraction 1 is shown in Figure 2.

Judul

APLIKASI PENJUALAN BARANG PADA TOKO ITSARO
DENGAN NETBEANS 5.5 DAN MYSQL

Abstraksi

Proses transaksi dalam kegiatan penjualan yang cepat dan tepat merupakan faktor yang sangat penting dalam pengoperasian suatu perusahaan. Sehingga diperlukan suatu aplikasi penjualan yang mampu mengolah transaksi secara efektif dan efisien agar dapat meningkatkan pelayanan terhadap pelanggan atau konsumen dan peningkatan pendapatan perusahaan. Dalam penulisan ini penulis membuat aplikasi penjualan menggunakan Netbeans 5.5 dan MYSQL server 5.0 untuk pengolahan basis datanya. Pada aplikasi ini terdapat menu pelanggan, menu barang dan menu transaksi.

Figure 2. Abstraction Document Modification 1

3. Abstraction Document Modification 2

This modification of abstraction document is the document abstraction modification of the original document. The modifications to be done is change the position of the sentence contained in the original document abstraction. The example of abstraction document modifications 2 is shown in Figure 3.

Judul

APLIKASI PENJUALAN BARANG PADA TOKO ITSARO
DENGAN NETBEANS 5.5 DAN MYSQL

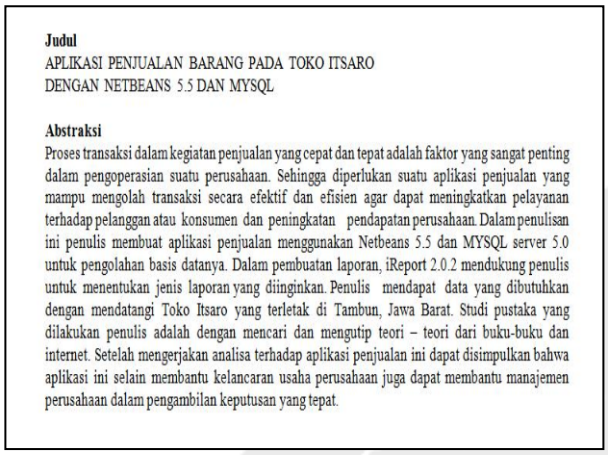
Abstraksi

Dalam penulisan ini penulis membuat aplikasi penjualan menggunakan Netbeans 5.5 dan MYSQL server 5.0 untuk pengolahan basis datanya. Dalam pembuatan laporan, iReport 2.0.2 mendukung penulis untuk menentukan jenis laporan yang diinginkan. Proses transaksi dalam kegiatan penjualan yang cepat dan tepat merupakan faktor yang sangat penting dalam pengoperasian suatu perusahaan. Sehingga diperlukan suatu aplikasi penjualan yang mampu mengolah transaksi secara efektif dan efisien agar dapat meningkatkan pelayanan terhadap pelanggan atau konsumen dan peningkatan pendapatan perusahaan. Penulis mendapat data yang dibutuhkan dengan mengunjungi Toko Itsaro yang terletak di Tambun, Jawa Barat. Studi pustaka yang dilakukan penulis adalah dengan mencari dan mengutip teori – teori dari buku-buku dan internet. Setelah melakukan analisa terhadap aplikasi penjualan ini dapat disimpulkan bahwa aplikasi ini selain membantu kelancaran usaha perusahaan juga dapat membantu manajemen perusahaan dalam pengambilan keputusan yang tepat.

Figure 3. Abstraction Document Modification 2

4. Abstraction Document Modification 3

Abstraction document modification 3 is modification of the original abstract document. The modifications to be done is change a few words with synonyms. In this Abstraction document modification 3, the word in the first sentence, namely "*merupakan*" is replaced by it's synonyms. The synonym word is "adalah". The second word which is replaced is "mengunjungi". That word is replaced by "mendatangi" and the third word, namely "melakukan" to be replaced with the word of "mengerjakan". Examples of abstraction document as seen in the Figure 4.



Judul
APLIKASI PENJUALAN BARANG PADA TOKO ITSARO
DENGAN NETBEANS 5.5 DAN MYSQL

Abstraksi
Proses transaksi dalam kegiatan penjualan yang cepat dan tepat adalah faktor yang sangat penting dalam pengoperasian suatu perusahaan. Sehingga diperlukan suatu aplikasi penjualan yang mampu mengolah transaksi secara efektif dan efisien agar dapat meningkatkan pelayanan terhadap pelanggan atau konsumen dan peningkatan pendapatan perusahaan. Dalam penulisan ini penulis membuat aplikasi penjualan menggunakan Netbeans 5.5 dan MYSQL server 5.0 untuk pengolahan basis datanya. Dalam pembuatan laporan, iReport 2.0.2 mendukung penulis untuk menentukan jenis laporan yang diinginkan. Penulis mendapat data yang dibutuhkan dengan mendatangi Toko Itsaro yang terletak di Tambun, Jawa Barat. Studi pustaka yang dilakukan penulis adalah dengan mencari dan mengutip teori – teori dari buku-buku dan internet. Setelah mengerjakan analisa terhadap aplikasi penjualan ini dapat disimpulkan bahwa aplikasi ini selain membantu kelancaran usaha perusahaan juga dapat membantu manajemen perusahaan dalam pengambilan keputusan yang tepat.

Figure 4. Abstraction Document Modification 3

4.2 Testing Results

The testing tobe performed are as follows:

1. Testing the abstraction of the original document with the three abstraction documents that we have been described above.
2. Testing these documents by using the four methods there are the key word, Karp Rabin, Jaro-Winkler Distance and manual methods.

The testing results that have been made to the three documents can be seen in Table 1.

Table 1. Comparison of Indonesian Language Documents Similarity Research Results

Document	Method	Abstraction Document Modification 1	Abstraction Document Modification 2	Abstraction Document Modification 3
Abstraction Document 1	Keyword	60,00	100,00	100,00
	Karp Rabin	50,00	83,30	75,00
	Jaro Winkler Distance	51,00	53,00	97,40
	Manual	50,00	100,00	100,00
Abstraction Document 2	Keyword	77,77	100,00	100,00
	Karp Rabin	66,66	50,00	50,00
	Jaro Winkler Distance	46,20	53,80	96,20
	Manual	66,66	100,00	100,00
Abstraction Document 3	Keyword	75,00	100,00	100,00
	Karp Rabin	60,00	60,00	20,00
	Jaro Winkler Distance	71,40	71,40	78,60
	Manual	66,66	100,00	100,00

5. ANALYSIS OF TESTING RESULTS

From the testing results that has been done, it can take some analysis, namely:

1. To compute the similarity of documents, can be used two ways, calculation of the number of the same word and the number of the same sentence. In the method of TF / IDF and the Jaro Winkler Distance determination of similarity is calculated based on the number of the same words, while in the Rabin and Karp's method and manual method, the analysis process to be done with calculate the same sentence.
2. Calculating the similarity using the keyword method got results close to 100%. This is because the keywords that generated the original document is only a few keywords. Keywords that generated just the words that includes words in programming languages, common words in computer science, so that not all words are compared.
3. Calculating the similarity using Karp Rabin method got almost the same results with the manual except for the third document which was modified by the word synonyms. This is because, Rabin Karp's method assumes that words that compared different so that a different meaning.
4. Calculating the similarity using Jaro Winkler Distance method did not get the same results with a manual, because the word to be compared is the same word.
5. The original document is compared with the first document that has been modified. The document is modified by simply taking part of the sentence. The results are obtained from Rabin Karp method is similar to the results are obtained using manual method.
6. Original document is compared with the second document that has been modified. The documen is modified in such way that the position of sentence is moved. Results obtained by the keyword method produces the same results with manually.
7. The original document is compared with the third document that have been modified in such way that some words replaced by synonyms of the word. The main result is the keyword method produces the same results with manually. This is because the comparison just keywords.

6. CONCLUSION

From the testing that has been done, the best method is Karp Rabin method, except for a synonymous

7. REFERENCES

[1] Ana Kurniawati, Lily Wulandari dan I Wayan Simri Wicaksana, *Perbandingan Tools Deteksi Plagiarisme untuk Dokumen*, Seminar Riset dan Teknologi Informasi, STMIK Akakom Yogyakarta, 8 Agustus 2009.

[2] Ana Kurniawati dan I Wayan Simri Wicaksana, *Perbandingan Pendekatan Deteksi Plagiarism Dokumen Dalam Bahasa Inggris*, Seminar Ilmiah Nasional, Universitas Gunadarma, Jakarta, 20-21 Agustus 2008.

[3] Benno Stein and Sven Meyer zu Eissen, 2006, *Near Similarity Search and Plagiarism Analysis*, Conference of

German Classification Society Magdeburg, ISBN 1431-8814, pp. 430-437.
[4] Felicia Utorodewo, 2007, *Bahasa Indonesia : Sebuah Pengantar Penulisan Ilmiah*, Penerbit Fakultas Ekonomi Universitas Indonesia.
[5] Junaiyah H Matanggui, 2009, Kamus Sinonim, Penerbit PT Gramedia Widiasarana Indonesia, Jakarta.
[6] Kamus Besar Bahasa Indonesia, 1997, Pusat Bahasa Departemen Pendidikan Nasional, Jakarta.
[7] Manber, 1994, *Finding similar files in a large file system*, Winter USENIX Technical Conference 1994, San Francisco, CA, USA.
[8] Mate Pataki, 2003, *Plagiarism Detection and Document Chunking Methods*, ACM.
[9] Parvati Iyer and Abdipsita Singh, 2005, *Document Similarity Analysis for a plagiarism detection system, 2nd Indian International Conference on Artificial Intelligence (*IICAI-05), pp. 2534-2544.
[10] Peter Vamplew and Julian Dermaoudly, 2005, *An Anti-Plagiarism Editor for Software Development Courses*, Conferences in Research and Practise in information Technology, Vol 42, Australia.
[11] Sinta Agustina, 2008, *Aplikasi Anti Plagiatisme Dengan Algoritma Karp-Rabin Pada Penulisan Ilmiah Universitas Gunadarma*, Jakarta.
[12] Sazali Rahman, 2009, *Aplikasi Perbandingan Kesamaan Antar Dokumen Dengan Algoritma Jaro-Winkler Distance Sebagai Deteksi Plagiarisme Pada Penulisan Ilmiah Universitas Gunadarma*, Jakarta.

The Effects of Training Documents, Stemming, and Query Expansion in Automated Essay Scoring for Indonesian Languange with VSM and LSA Methods

Heninggar Septiantri
Faculty of Computer Science
Universitas Indonesia, Depok, Indonesia
hes51@ui.ac.id

Indra Budi
Faculty of Computer Science
Universitas Indonesia, Depok, Indonesia
indra@cs.ui.ac.id

ABSTRACT
Research in automated essay scoring system has been done using Latent Semantic Analysis (LSA) method. One of the limitations is the lack of training documents to optimize LSA results. Regarding such limitation, the use of Vector Space Model (VSM) can be considered. This research aims to compare LSA and VSM to score essay answer. Experiments are done with 13 problems with 42 test participants. Overall results show that average correlation of score between VSM-human is higher than LSA-human.

Keywords
Automated essay scoring system, Latent Semantic Analysis, Vector Space Model, stemming, query expansion.

1. INTRODUCTION
Research in automatic essay scoring has been started since 15 years ago through a research conducted by Page, resulting in a system called PEG (Project Essay Grader). PEG scoring the writing styles or techniques by measuring intrinsic factors such as the essay length, diction, etc. Later on, essay scoring system was evolving, scoring not only writing technique but also the content. Various methods are used, from statistical methods to Natural Language Processing (NLP). The example of essay scoring system including IEA (Intelligent Essay Assessor), E-Rater, and C-Rater [1].

Starting from 2005, researches in automated essay scoring for Indonesian language has been conducted in Indonesia. One of them resulting a system named SIMPLE, dedicated for scoring essay in Indonesian language, and was developed at The Electrical Engineering Department, Faculty of Engineering University of Indonesia. SIMPLE used statistical technique namely Latent Semantic Analysis (LSA). It was the same method used by IEA.

Table 1 describes the summary of the researches conducted regarding SIMPLE.

Table 1. Prior Researches in Automated Essay Scoring System for Indonesian Language (SIMPLE)

Research	Results
Brian Prima Krisnanda [2]: using 10, 20, 30, and 40 keywords	Correlation between scores yielded by system and scores given by human was 0.86-0.96. More keywords resulting in higher correlation.
Ratna, Budiardjo, Hartanto [3]: addition of term wight, word order, and word similarity	Agreement between human scores and system was 69.80-94.64% (experiment with 5 students), and 77.18%-98.42% (experiment with 10 students).
Dudi Hermawandi [4]: implementation of SICBI (Sqrt-IGFF-Cosn-Bnry-IDFB) weighting scheme.	Average scores differences between system and human was 13.98, 17.84, dan 10.90, in experiment with 10, 15, and 20 students (consecutively) for 10-question examination.
Diego Octaria [5]: implementation of 4 weighting schemes	The highest correlation between system's scores and human's scores was 0.39, in experiment with 20 students, scored using Sqrt-Normal-Cosn-Bnry-Normal weighting sheme.
Harisma [6]: implementation of 3 keyword weighting scheme (weighted 1, 2, and 3)	Correlation between system's scores and human's scores was 0.77 and average scores difference was 17.36, in experiment with 10 essay questions for 23 students.

Those researches were only using student's essays and key answers to build the LSA semantic space. Whereas LSA needed large-scale training document to build the semantic space so that the term similarity could be found [7]. The lack of training document will result in not optimal results and the system will not be able to recognize word similarities, as stated by Octaria [5].

With limited training documents, LSA will not be able to give optimal results. In such condition, the use of VSM (Vector Space

Model) could be considered. One of VSM's drawbacks is that it can't recognize word similarities. To overcome that situation, we need additional techniques in order to make VSM recognizes word similarities. The techniques that can be used are query expansion and stemming (affix removal from words).

This research is aimed to compare the effectiveness of automated essay scoring using LSA and VSM, and to examine the effects of query expansion, stemming, and training documents to the system's results.

2. THEORETICAL BACKGROUND

2.1. Vector Space Model (VSM)

VSM is a representation of a collection of documents as vectors in a vector space. VSM is a basic technique in information retrieval that can be used to assess the relevance of retrieved documents to the keyword search (query) on search engines, classification of documents, and clustering of documents [8].

Examples of changes from text to the vector representation is as follows:

There are two documents and a query:

Doc1: "morning"
Doc2: "this sunny morning"
query: "this morning"

Word/term appearance in each documents are counted. Doc1, Doc2, and query are vectors representing the initial text.

Table 2. Vectors Representing Texts

term	*Doc1*	*Doc2*	*query*
morning	1	1	1
this	0	1	1
sunny	0	1	0

The visualization of the representation is depicted in Figure 1.

The closer the distance between vectors of documents/queries, the more similar the content. In Figure 1, document 2 (Doc2) is more similar to the query. The calculation of similarities between documents is done using cosine similarity.

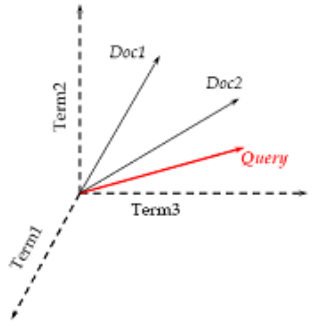


Figure 1. Vector Space Model

2.2. Latent Semantic Analysis (LSA)

LSA is a method to determine the similarities between word meaning or document content by analyzing large-scale text corpus [7]. LSA is not using NLP processing of artificial intelligence program, but only using pure mathematical/statistical methods which can extract and infer relationship between term in a document according to the contextual usage.

The beginning of the LSA process is the same with VSM, which is representing the text to the vector. It's just that these vectors then combined into a matrix A which then decomposed into three matrices components (U, S, V) through Singular Value Decomposition process (SVD).

$$A_{mxn} = U_{mxm}\, S_{mxn}\, V^{T}{}_{nxn}$$

Three components of the matrix consists of a row orthogonal matrix, a column orthogonal matrix, and a diagonal matrix. The resulting diagonal matrix containing the nonnegative elements, and a non-zero elements are called singular values of A.

From this matrix decomposition, dimension of matrix A can be reduced to the size of k and the matrix A can be reconstructed using only k dimension to approximate matrix A. Reconstruction with the reduced dimension produces matrix Ak, which is the so-called low-rank approximation of matrix A [8]. Reconstruction with only k dimension is done by taking only k dimension from matrix components U, S, and V so that $A_k = U_k S_k V_k{}^T$.

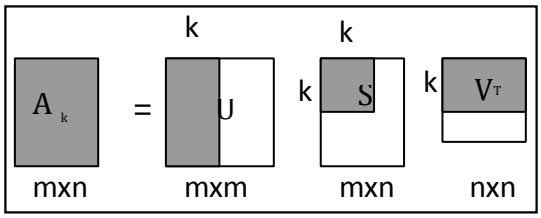


Figure 2. SVD Process with Reduced Dimension

Low rank approximation of A produces a new representation for each document, in which the similarity between the words and documents in the matrix A can be discovered. The rank used in SVD will be different for each case. If the rank is too small then the estimation of the similarity between terms/documents will be too high. Conversely, if the rank is too high, the similarity between terms/documents can not be caught. Rank selection can be done by trying some of the commonly used rank and selecting rank that gives optimal results.

2.3. Query Expansion and Stemming

Expansion of the key answer or query expansion is a technique used on search engines where users insert additional words in the keyword search (query). This is carried out to improve the recall, or the ratio between the number of relevant documents that is retrieved with the total number of all relevant documents [8].

Stemming or affix removal is commonly used technique in information retrieval process to omit the morphological variations. Stemming program or stemmer usually consists of a set of rules and dictionary [9].

Stemming is used because there are some different form of words with the same meaning, for example democratic, democracy, and

democratization (or "demokratik", "demokrasi", and "demokratisasi" in Indonesian). In some situations, it will be beneficial if the document searching with a keyword/query resulting in documents that containing that keyword in various forms [8].

3. DESIGN AND IMPLEMENTATION

Automated essay scoring receives input in the form of students' answers and answer key from teacher. Collection of students' answers are represented in the term-document matrix and the answer key is represented in answer vector. The similarity between the student's answers with an answer key is yielded from the value of cosine similarity between students' answers vector (each column of the term-document matrix) with the answer key vector. This value is then used as a student's score. Figure 3 describes the process flow of the system.

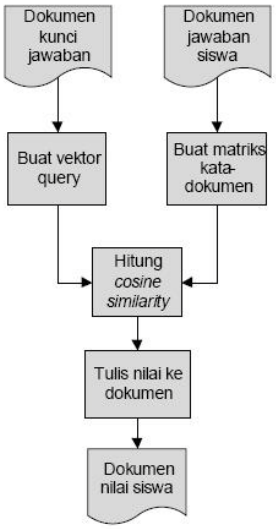


Figure 3. Process Flow of Automated Essay Scoring System

Generally, LSA system and VSM system proceeds with this flow. The difference is that in LSA system, after building term-document matrix, SVD is applied to the matrix. After that, the cosine similarity between student answer vector (each column of the matrix) and key answer vector is calculated.

This research adds the use of query expansion and stemming. Both techniques are applied to the input document before it is represented as a matrix. In addition, training document is also used in the LSA scoring system. In LSA scoring, various combinations are used to form the semantic space and the asnwer key vector. The combinations are described in the following table.

Table 3. Combinations of LSA Scoring Schemes

LSA Scheme	Semantic Space	Query
LSA1	Collection of students answers	Answer key
LSA2	Training documents and answer key	Collection of students answers
LSA3	Training documents and collection of students answers	Answer key

The system is implemented in Matlab and Perl programming language. In addition, it is also using Matlab Tool TMG (Text to matrix Generator) [10] to create term-document matrix and query vector. Another program that is used outside the system is stemmer for Indonesian language which is developed in Java programming language [9].

Overall, the system is composed in six programs as described in Figure 4.

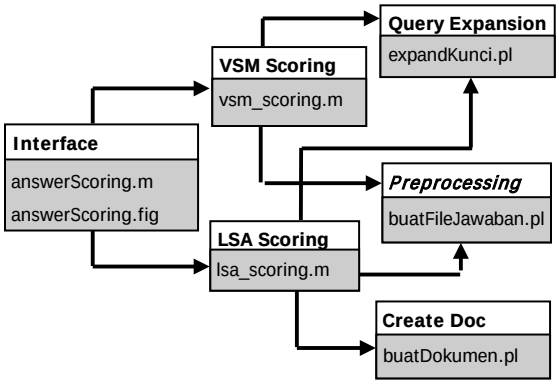


Figure 4. The Implementation Structure of The System

3.1. System's Interface

This program is designed to bridge the users with the system. Through this interface, users can choose several options related to automatic essay scoring system as follows:

1. Options of term weighting scheme for students anwers and answer key.
2. Options of scoring methods, consists of VSM, LSA1, LSA2, and LSA3.
3. Option to use query expansion
4. Option to use stemmer for English language (Porter Stemmer) which is integrated with the Matlab Tool TMG.

3.2. Query Expansion

This program reads the initial answer key document and opens the document containing synonym list, and then compare the keyword in asnwer key with the words on the list. If there is a synonym (or synonyms) for a particular word in the answer key, then all these synonyms are added to the answer key.

3.3. Preprocessing

Before the scoring process is carried out, input documents are first preprocessed. Particular parts in input documents are marked with tags to ease the text processing. The example of input documents are given in Figure 5.

The information included within each tags are taken. Each problem could consists of one or more subproblem. The division of problem into subproblems is aimed to make the context in the answer more specific. Collection of all students answers for the same subproblems is written in a file, one file for one subproblem. Answer key for each subproblem is also written to different file.

3.4. VSM Scoring Program

This program reads answer key document and students answers document for each subproblem and then creates term-document matrix (matrix A) from students answers and query vector (vector Q) from answer key. A and Q is made by using tmg_query function from TMG. The visualization of matrix A and vector Q is depicted in Figure 6.

student answer

<mhs>1</mhs>
<sub>
strategy suatu formulasi yang dirancang sehingga perusahaan dapat bertahan dan mencapai tujuannya biasanya dilengkapi dengan perencanaan dan aturan2
</sub>
<sub>
a forecasting => memprediksi bisnis teknologi ekonomi trend dll
b resource allocation => mengalokasikan sumber daya yang dibutuhkan
c core competency => bisnis utama yang dijalankan
</sub>

answer key

<subsoal>2</subsoal>
<sub>
strategi formula menyeluruh bisnis bersaing tujuan rencana kebijakan diperlukan untuk mencapai tujuan goal
</sub>
<sub>
perkiraan forecasting
alokasi sumber daya resource allocation
kompetensi inti core competence
analisis lingkungan industri industry environment analysis
analisis perusahaan company analysis
rencana bisnis business plan
</sub>

Figure 5. Input Documents Example

3.5. LSA Scoring Program

Before creating the matrix and vector, this program creates documents by executing the program buatDokumen.pl. This program only served to unite the input documents (students answers, answer key, and training documents) into one document.

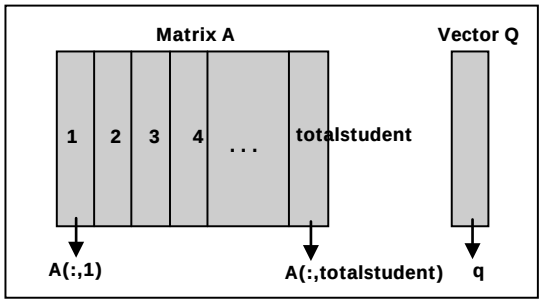


Figure 6. Matrix A and Vector Q for VSM Scoring

The documents created will vary according to the selected marking scheme (stated in Table 3). LSA1 only require students answers and answer key, while LSA2 and LSA3 require additional training documents.

The documents required and the visualization of generated matrix A and vector Q for LSA1 is depicted in Figure 7.

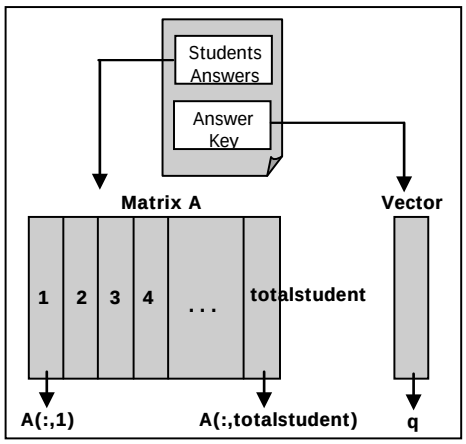


Figure 7. Documents, Matrix A, and Vector Q for LSA1

LSA2 scheme is using additional training documents to form the term-document matrix. LSA3 is also using training documents, but what makes it different from LSA2 is that LSA3 uses training documents and students answers to form the term-document matrix, and answer key to form the query vector, while LSA2 uses training documents and answer key to form the matrix, and students answers as vectors.

4. EXPERIMENTS

4.1. Experiments Environment

Experiments were undertaken on a computer with an Intel Celeron M processor 1.73 GHz, 1 GB DDR2 memory, and 80 GB harddisk. The required softwares are Matlab, Matlab Tool Text to Matrix Generator (TMG), Perl, and Java Standard Development Kit JDK1.6.0_03. The operating system used is Windows XP.

4.2. Experiments Documents

Experiments were conducted with 546 essay answers consists of 13 questions/problems, answered by 42 students for each question. Essay answers were taken from End-Term Examination of E-Commerce course in the Faculty of Computer Science Universitas Indonesia in 2008.

The examples of questions given in the exam are the following.

Table 4. Essay Problems for Experiments

Problem	Question
2	Sebutkan lima atribut *m-commerce*
3	Bandingkan antara isu legal dan isu etika
5	Apakah yang dimaksud dengan P2P *payments*? Berikan dua contohnya

9	Definisikan strategi dan sebutkan tiga elemen strategi
10	Sebutkan tiga fungsi utama pasar
11	Apa yang dimaksud dengan segmentasi pasar? Bagaimana menggunakan internet untuk melakukan segmentasi pasar?
12	Jelaskan apa saja jenis *cyber crime* yang sudah diatur dalam UU ITE

4.3. Experiments Details

Below are the detail of experiments conducted:

1. Experiment 1, using students answers and answer key without any addition.
2. Experiment 2, using students answers and expanded answer key.
3. Experiment 3, using students answers and answer key, both have been processed with stemmer.
4. Experiment 4, using students answers and expanded answer key, both have been processed with stemmer.

Those four experiments are applied for VSM and LSA scoring scheme (LSA1, LSA2, and LSA3).

Constant variables used in all experiments are the following:

1. Weighting scheme for students answers and answer key is local weighting logarithm. This weighting scheme is used because it gives proportional weight to words so that the words that frequently appear will not be given high weight.
2. Stopwords removal for Indonesian language. Stopwords are commonly-used words that often appear in documents, such as “and”, “with”, “or”, etc.
3. The rank used in LSA1 is 2 (k=2), in accordance with prior research by Hermawandi, Octaria, and Harisma. For LSA2 and LSA3, the rank used is 60. This rank was chosen because it gave best result in prior experiment with one essay problem. Prior experiment initially conducted with ¼, ½, and $^1/_8$ of matrix dimension of the training documents (dimension 200). The best result was obtained with rank 50, and next experiment with rank 60 yielded better result.

5. RESULTS AND DISCUSSION

This section describes all of the experiments results and discussion.

5.1. Experiment 1

Overall, the average correlation between the scores given by human (HR-Human Rater) and VSM system is greater than the average correlation between scores given by human and system LSA1, LSA2, and LSA3. This may occur because of the lack of training documents to build the LSA semantic space so that the LSA cannot give optimal results. Although training documents have been used in LSA2 and LSA3, in average the correlation between scores given by LSA system and human is still lower than of VSM. This might be caused by small number of training documents and the contents of the documents are less specific for each problems domain. The correlation between VSM, LSA1, LSA2, and LSA3 for all problems are visualized in Figure 8.

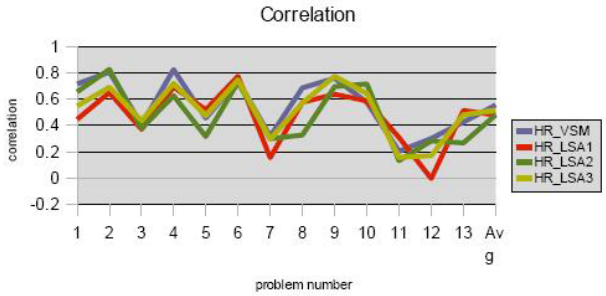


Figure 8. Correlation between HR and VSM-LSA (Exp1)

5.2. Experiment 2

The influence of the answer key/query expansion in VSM system is reducing the average correlation between system's scores and human rater, from 0.56 to 0.55. While there is an increase in average correlation between LSA1 system's scores with human rater, from 0.48 to 0.49. LSA2 system also undergo an increase of the average correlation with the human rater, from 0.48 to 0.49. So does LSA3 who experienced an increase in average correlation with the human rater from 0.52 to 0.53. A graph depicting the correlation between human rater and VSM, LSA1, LSA2, and LSA3 in experiment 2 is given in Figure 9.

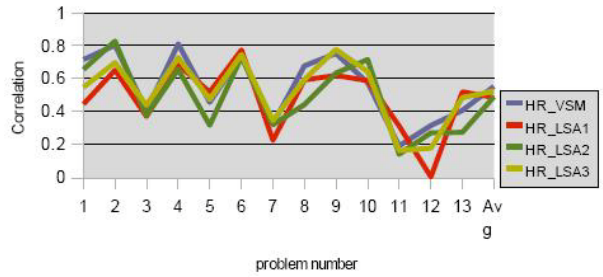


Figure 9. Correlation between HR and VSM-LSA (Exp2)

Answer key/query expansion can increase the possibility of matching between varying words used in answers key and words used by studentsr. In VSM systems, the query expansion is reducing the correlation precisely. This may occur because many synonyms are included as a keyword, while probably no words are match with the student's used words. Answer key vector length becomes longer so that it can lower the score obtained by student. Unlike the VSM, the query expansion in LSA system increases the correlation. This is probably because the addition of an appropriate synonym in the context of the answer key could clarify the context of the answer key and student answers, so LSA could perform better similarity analysis.

5.3. Experiment 3

Affix removal in VSM increasing it's average correlation of scores with human rater. Conversely, in LSA1 and LSA3 it is decreasing the average correlation, while in LSA2 the average correlation remain constant. Affix removal causing more words can be matched because it produces the word base. In the VSM, it is beneficial because VSM is only judging similarities by

matching the answer key words to students' answers. Whereas in LSA which analyzes the emergence of words not only in just one answer, but from all the answers, this is not beneficial. Because the use of affix removal could reduce the context of the answers/ key thereby reducing the quality of analysis. This influence is difficult to observe directly because LSA is inferring similarities by analyzing the appearance of the words globally in whole documents. A graph depicting the correlation between human rater and VSM, LSA1, LSA2, and LSA3 in experiment 3 is given in Figure 10.

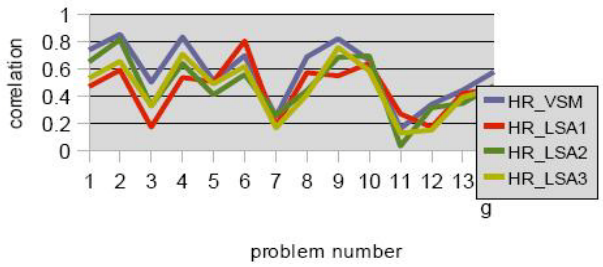


Figure 10. Correlation between HR and VSM-LSA (Exp3)

5.4. Experiment 4

The use of combination between query expansion and stemming gave different results. The average correlation with human rater of VSM and LSA2 scores are increasing, while LSA1 remain constant, and LSA3 decreasing. Probably a greater influence (more dominant) is given by the stemming because changes in affix removal occurs in the entire document, while the expansion of the answer key may be less influencing because only affects in adding synonyms for 29 words. A graph depicting the correlation between human rater and VSM, LSA1, LSA2, and LSA3 in experiment 4 is given in Figure 11.

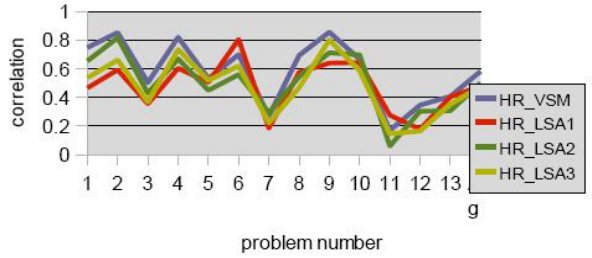


Figure 11. Correlation between HR and VSM-LSA (Exp4)

6. CONCLUSION

The conclusion of this research are the following:

1. Overall, the average correlation of scores between human rater and VSM system are higher than of LSA. This might be due to the lack of training documents to build enough LSA semantic space. With limited document, the use of VSM may be an alternative for assessing essay exam answers.
2. The use of stemming in VSM system has increased the average correlation of scores between human rater and system, while in LSA it is reducing the average correlation. This is probably because the VSM that depends on word matching could caught more similaritis due to affix removal from words. While LSA that used the global analysis of the appearance of words on the answer may experience changes in interpretation caused by lack of clarity in the context oduw to affix removal
3. The use of query expansion in VSM system slightly lessen the average correlation between system and human rater, while in LSA it is increasing the average correlation. This is probably because the addition of synonyms could clarify the context of a word in a passage so that LSA can capture the relationship between words/documents better VSM.
4. The use of combination between stemming and query expansion of the answer keys is inconclusive because the effects were vary. The average correlation with human rater of VSM and LSA2 scores are increasing, while LSA1 remain constant, and LSA3 decreasing. Probably a greater influence (more dominant) is given by the stemming because changes in affix removal occurs in the entire document, while the expansion of the answer key may be less influencing because only affects in adding synonyms for 29 words.

7. REFERENCES

[1] Valenti, S., Neri, F., Cucchiarelli, A. (2003). An overview of current research on automated essay grading. *Journal of Information Technology Education, Volume 2.*

[2] Krisnanda, B. P., (2005). Sistem penilaian essay otomatis dengan menggunakan metode LSA. Depok: Fakultas Teknik Universitas Indonesia.

[3] Ratna, A.A.P., Budiardjo, B., Hartanto, D. (2007, April). SIMPLE: sistim penilai esei otomatis untuk menilai ujian dalam bahasa Indonesia. *Jurnal Makara Teknologi*, 5-11.

[4] Hermawandi, D. (2008). Implementasi pembobotan SICBI pada aplikasi essay grading metode LSA. Depok: Fakultas Teknik Universitas Indonesia.

[5] Octaria, D. (2008). Implementasi skema pembobotan pada aplikasi penilaian esai otomatis metode LSA. Depok: Fakultas Teknik Universitas Indonesia.

[6] Harisma, N. Z. (2008). Implementasi sistem penilaian esai otomatis metode LSA dengan tiga bobot kata kunci. Depok: Fakultas Teknik Universitas Indonesia.

[7] Landauer, T. K., Foltz, P. W., Laham, D. (1998). Introduction to latent semantic analysis.

[8] Manning, C. D., Raghavan, P., Schutze, H. (2008). *Introduction to information retrieval*. New York: Cambridge University Press.

[9] Adriani, Mirna (2008). *Information retrieval*. Modul kuliah Pemrosesan Teks Fakultas Ilmu Komputer UI semester ganjil 2009.

[10] Zeimpekis, D., Gallopoulos, E. (2008). Text to matrix generator user's guide.

http://scgroup.hpclab.ceid.upatras.gr/scgroup/Projects/TMG

The Impact of Object Ordering in Memory on Java Application Performance

Amil A. Ilham
Institute of Systems, Information Technologies and Nanotechnologies
2-1-22, Momochihama, Sawara-ku, Fukuoka, Japan
+81-92-852-3460
amil@isit.or.jp

Kazuaki Murakami
Institute of Systems, Information Technologies and Nanotechnologies
2-1-22, Momochihama, Sawara-ku, Fukuoka, Japan
+81-92-852-3460
murakami@it.kyushu-u.ac.jp

ABSTRACT
Java is gaining popularity in software development. It is widely used in network computing and embedded systems because it offers several key advantages such as safe programming, code verification and checking, automatic memory management, and significant support from the computing industry. This paper is aimed to evaluate the impact of different object orders in memory on Java application performance. This work is motivated by the facts that Java programs create many objects dynamically on the heap but never freed explicitly by the code. A Java Virtual Machine (JVM) implements a garbage collector to automatically collect objects that are no longer accessed by the program and to make the space available for new object allocations when the heap is full. Once the garbage collection is finished, the live objects remain in the heap. These objects might be spread across the memory since they are not necessarily resided in adjacent memory locations. If the objects are compacted, their ordering might not match with the traversal order of the program. This means that the remaining live objects in memory after garbage collection are sensitive to cache misses and TLB misses and the application execution time might suffer from the penalty of poor spatial locality of objects in memory. To show how the order of objects in memory affects Java application performance, we implemented two different copying order schemes at garbage collection time: Bread First (BF) scheme and Depth First (DF) scheme. Our experiment results show that Java execution time, cache misses and DTLB misses vary by 3-16%, 5-20% and 9-21% respectively due to BF and DF schemes.

Keywords
Java, object, memory, garbage collection, cache

1. INTRODUCTION
The clear software engineering and security advantages of Java programs have attract programmers to use this language to write all kinds of applications. Automatic memory management in Java increases the productivity of programmers by reducing programmer burden and eliminating sources of errors [13]. As the popularity of Java has been increased among programmers, researchers have been paid more attention to improve the performance of Java programs.

This paper is aimed to evaluate the impact of different object orders in memory on Java application performance. Understanding object behaviors in Java is important because Java is object oriented program which creates many objects dynamically on a heap. During runtime, Java program will access and mutate the objects. When the order of objects in memory is not match with the way the Java program accesses the objects, the program performance might be degraded.

This work exploits the existence of garbage collector in Java Virtual Machine (JVM). Java creates many objects dynamically on a heap similar to the other program languages such as C and C++. However unlike C and C++, object deletion in Java is never done manually by programmers. JVM implements garbage collector to free unused objects (*referred to as garbage*). JVM halts the running application and invokes a garbage collector when no more space available in memory for new object allocation. Once the garbage collection is finished, the running application is resumed.

Garbage collection performs two distinct functions: distinguishing the live objects from the garbage in some way (*garbage detection*) and reclaiming the garbage objects' storage, so that the running program can use it (*garbage reclamation*). Garbage collector uses approximate liveness by reachability from outside the heap to detect garbage. Any object the program cannot reach is considered as garbage because a program can use only the objects that it can find.

The garbage reclamation is an important phase in garbage collection because it sweeps the garbage to reclaim the space and keeps the live objects remain in memory. There are three ways of garbage reclamation: *sweep-to-free*, *compacting* and *evacuation*. *Sweep-to-free* sweeps all garbage and keeps the live objects in memory unchanged. *Compacting* is similar to *sweep-to-free* except that it compacts live objects to the end side of the memory. *Evacuation* sweeps all garbage and moves all live objects to a reserved space.

2. BACKGROUND
There are two main types of garbage collectors: non-moving garbage collector and moving garbage collector.

2.1 Non-moving garbage collector
An example of non-moving garbage collector is a *mark-sweep* garbage collector [12]. This algorithm implements *sweep-to-free* garbage reclamation. Figure 1 shows the process of garbage detection and garbage reclamation using *mark-sweep* garbage collector.

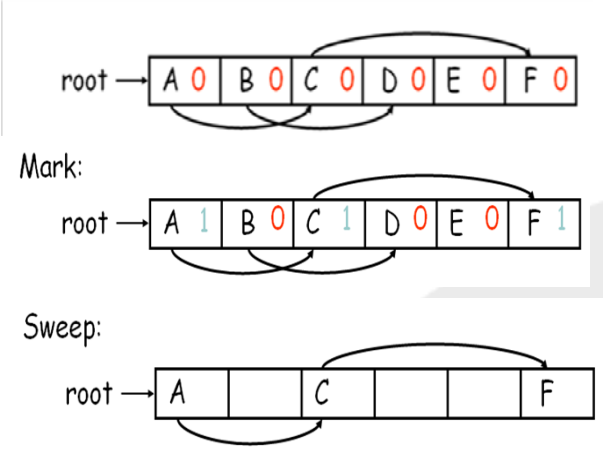


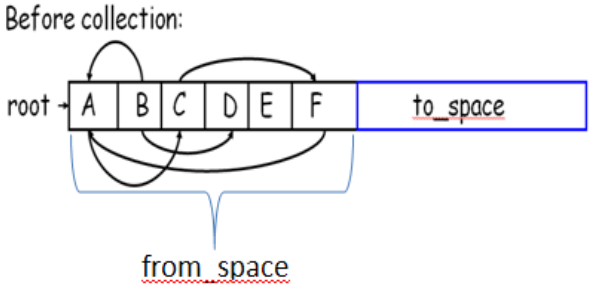
Figure 1. *Mark-sweep* garbage collector

As shown in Figure 1, garbage detection process is started by setting a marked bit 0 to all objects. The process is then continued to a marking phase which marks all live objects by changing their marked bits to 1. Objects are considered to be alive if they are reachable from the root or there are existing references from the root to the object. For example object A is a live object because there is a reference from the root to this object. Object F is also a live object because this object can be reached from the root through this path: root, reference to object A, object A, reference to object C, object C, reference to object F. On the other hand, object B is garbage because no reference exists between this object and the root. The last phase in *mark-sweep* garbage collection is the sweeping phase which sweeps all objects that have a marked bit 0. The live objects remain in the memory unchanged. Because this algorithm does not move live objects at garbage collection time, we cannot use it in our experiments.

2.2 Moving Garbage Collector

An example of moving garbage collector is a *semi-space* garbage collector. This algorithm implements *evacuation* garbage reclamation. Figure 2 shows the process of garbage detection and garbage reclamation using *semi-space* garbage collector.

This algorithm divides heap into two spaces: *from-space* and *to-space*. During the running application, only one space is used to allocate new objects. The other space is reserved and used as a place for object evacuation when the garbage collector is invoked.



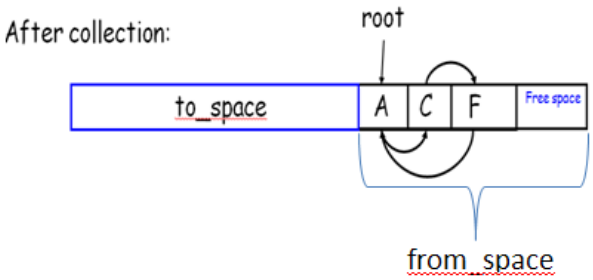


Figure 2. *Semi-space* garbage collector

As shown in Figure 2, during the running application (*before collection*), only the *from-space* is used for object allocation and the *to-space* is reserved. Once the *from-space* is full, JVM halts the running application and invokes the garbage collector to detect garbage and to reclaim the space. The garbage collector starts to find all live objects by inspecting the references to the object from the root. When the garbage collector finds a live object, it moves it to the reserved space, the *to-space*. For example, the garbage collector will move object A to the *to-space* because it has a reference from the root. Object C is also moved to the *to-space* because it has a reference from Object A which is a live object, and so on. After all live objects are evacuated to the *to-space*, JVM flips the role of the two spaces. The *to-space* becomes *from-space* and vice versa. The running application is resumed and the available space in the *from-space* is used for new object allocation (*after collection*).

This algorithm is less efficient in term of the use of memory space compare to *mark-sweep* garbage collector. However since it moves and compact the live objects, it improves object locality and eliminates defragmentation problems. For moderate to large heap size, the performance of *semi-space* garbage collector is better than the performance of mark-*sweep* garbage collector [5].

Semi-space garbage collector provides opportunity for object ordering evaluation through its garbage reclamation. During the reclamation process, live objects are copied and can be placed in any order in the reserved space. We use this garbage collector in our experiments and we implement two pre-determined object ordering schemes: *breadth first* scheme and *depth first* scheme. We then evaluate the impact of these two different schemes on the performance of Java applications.

2.3 Pre-determined object ordering

The properties of pre-determined object ordering are the order scheme is determined prior runtime and the order scheme is independent from the way the program accesses the objects at runtime. The order of object copying at reclamation time is performed based on the connectivity of the objects.

Figure 3 shows an example of object connectivity graph. This connectivity information is available at compilation time of Java applications. The important terms are *root*, *parent*, *child/children* and *siblings*. In this figure, object A is called root object, because there is no reference to this object. Object A has direct references to objects B, C, and D. These objects are called object A`s children and they are siblings. Similarly, object E and F are siblings and they are object B`s children.

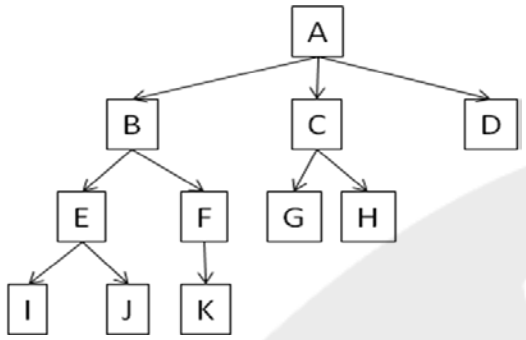


Figure 3. An example of object connectivity graph

Breadth first (BF) scheme and depth first (DF) scheme is copying objects based on this object connectivity information. The priority copying of BF scheme is all object`s siblings while DF scheme gives the priority copying to object`s children.

In more details, BF scheme would copy all children immediately after copying the parent. For example, in Figure 4, immediately after copying A, BF scheme would recursively copy all its children, object B, C and D. After all A`s children are copied, the next objects to be copied are all children of B, all children of C, and so on.

Unlike BF scheme, DF scheme would copy only one child immediately after copying the parent. For example, in Figure 4, after copying object A, DF scheme would recursively copy one of its children, say B, and the one of B`s children, say E, and so on.

The resulting object orders of BF scheme and DF scheme in memory are shown in Figure 4 and Figure 5 respectively.

Figure 4. Objet order in memory based on BF scheme

Figure 5. Objet order in memory based on DF scheme

3. EVALUATION ENVIRONMENT

3.1 Hardware and Operating System

We conducted our experiments on a single 3.20 GHz Pentium 4 with Hyperthreading disabled. It has a 64 byte DL1 and L2 cache line size, an 8KB 4-way set associative L1 data cache, a 512KB unified 8-way set associative L2 cache, and 2GB of main memory.

We performed the experiment on 32bit Linux 2.6.18 kernel with *perfctr* patch to access the Pentium 4`s on-chip performance counters to measure the number of L1 and L2 cache misses and DTLB misses. The computer run stand alone with all unnecessary daemons and services stopped. The network interface is also down.

3.2 Virtual Machine and Benchmarks

Our virtual machine infrastructure was Jikes RVM 3.1.0, released on June 10, 2009. The Jikes RVM is a performance-oriented, server-based, Java virtual machine from the IBM T.J. Watson Research Center [1][2]. The JVM was configured to compile all methods with optimizing compiler.

For each experiment, we run the JVM once and run each Java application twice. We measured the performance of Java application at the second run because Eeckhout et al. show that measurements of the first run of a Java application inside a JVM tend to be dominated by the JVM overheads instead of by application behavior [8].

We setup the JVM to invoke *semi-space* garbage collector with BF scheme or DF scheme and we run 11 Java applications from SPEC [14] and DaCapo [4] benchmark suite version 2006-10-MR2. We run all applications on 5 different heap sizes: minimum heap size (minHeap), 1.5x, 2x, 2.5x and 3x minHeap. For each heap size, we run the application 5 times and report the mean.

4. RESULTS

Before we present the impact of BF and DF schemes on Java application performance, we first report the minimum heap size for each application as shown in Table 1.

Table 1. Minimum heap size

Application	Description	minHeap (MB)
db	In-memory database	31
javac	Java compiler	36
jack	Parser generator	33
jess	Expert shell system	22
mtrt	Multi-threaded raytracer	34
raytrace	Raytracing	27
compress	Lempel-Ziv compressor	26
antlr	Parser generator	40
bloat	Bytecode optimizer	64
fop	XSL-FO to pdf converter	56
hsqldb	Database written in Java	184

The minimum heap size is the minimum memory required by the JVM to run the application without throwing out of memory errors. These numbers are obtained through trial and error experiments. To provide the same memory occupancy for each application in our experiments, we set the heap size as a multiplication of the minimum heap size, e.g. for 75% occupancy we set the heap size equals to 1.5x minHeap.

The next section shows measurements of the impact of BF and DF schemes on Java application performance.

4.1 Effect on execution time

Figure 6 shows the impact of BF and DF schemes on application execution time (1.5x minHeap). The horizontal axis shows the Java

applications and the vertical axis shows the execution time in ms (*left graph*) and the percentage difference in execution time when DF scheme is normalized to BF scheme (*right graph*).

The left graph shows that, in general, DF scheme improves execution time over BF scheme except for application compress. The impact is high for one application (db), moderate for eight applications (javac, jess, mtrt, jack, antlr, bloat, fop and hsqldb) and low for two applications (raytrace and compress). As shown in the right graph, the percentage difference in execution time for application db due to BF and DF schemes is around 16%. Applications which moderately affected by BF and DF schemes have execution time variation in the range of 3-7% while the other two applications have less than 2% differences in execution time.

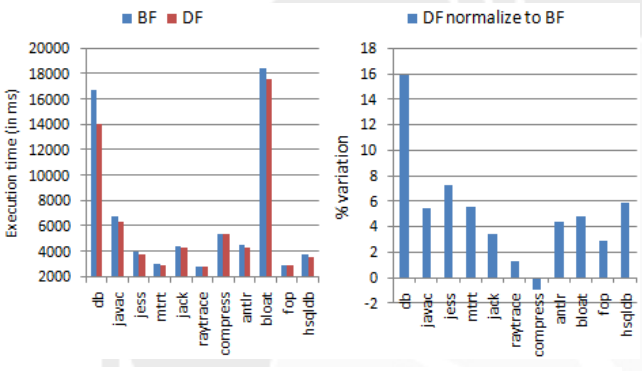
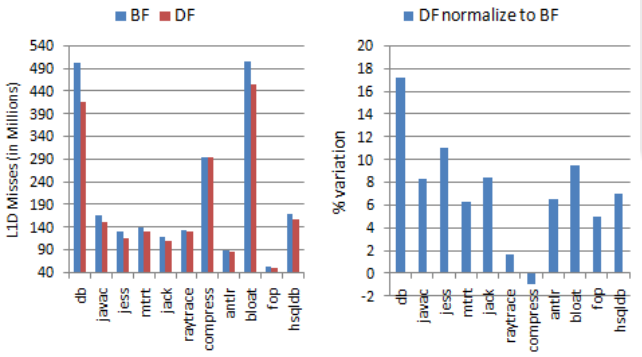


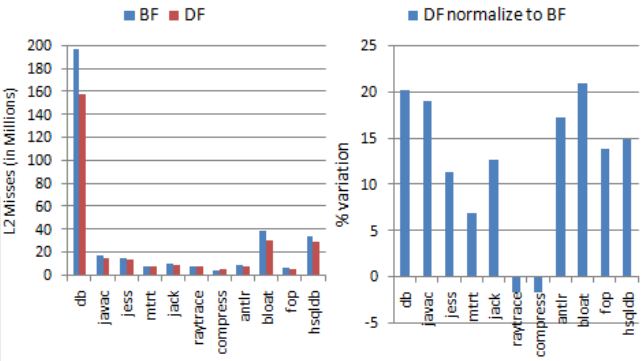
Figure 6. Total execution time

4.2 Effect on Cache and DTLB Misses

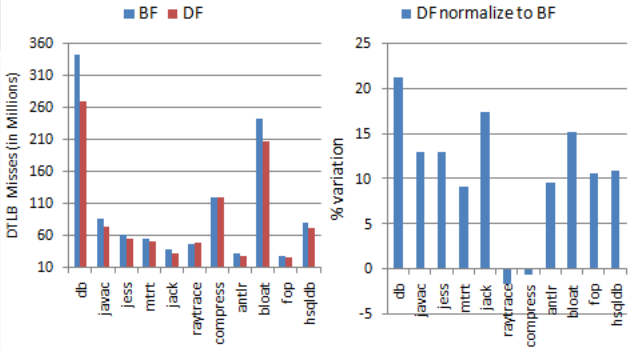
This section shows how BF and DF schemes affect cache and DTLB misses of the applications. Figure 7 shows the number of L1D cache, L2 cache and DTLB misses due to BF and DF schemes (1.5x minHeap). As expected, BF and DF schemes have a significant impact on cache and DTLB misses. The trend is similar to the effect on execution time but the percentage differences in cache and DTLB misses are larger than the percentage differences in time. As shown in Figure 7a, BF and DF schemes cause up to 17% differences in L1D cache misses for application db. Moderately affected applications have 5-11% differences in L1D cache misses due to BF and DF schemes. Raytrace and compress which tend to be insensitive to BF and DF schemes have less than 2% variations in L1D cache misses.



(a) L1D cache misses



(b) L2 cache misses

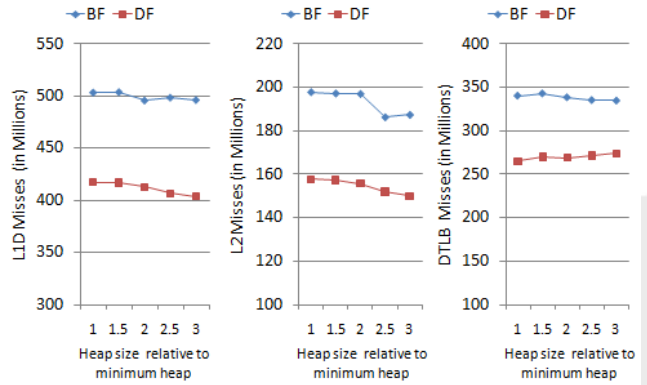


(c) DTLB misses

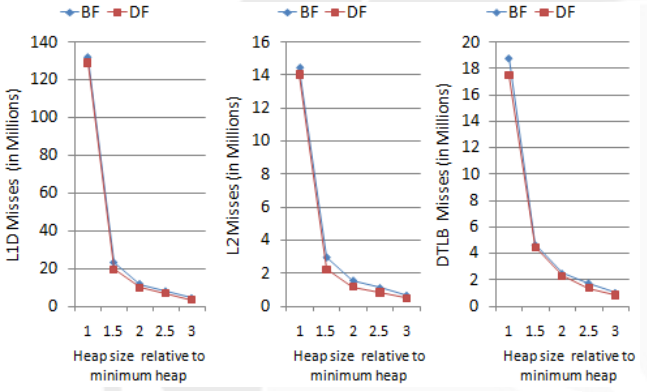
Figure 7. L1D cache, L2 cache and DTLB misses

BF and DF schemes more strongly affect L2 cache and DTLB misses compare to L1D cache misses. They cause 6-20% differences in L2 cache misses (Figure 7b) and 9-21% differences in DTLB misses (Figure 7c) for all applications except raytrace and compress. These two applications which just have less than 2% variations in L2 cache and DTLB misses are confirmed to be very less affected by object ordering BF and DF schemes.

Looking at individual application, DF scheme substantially improves performance of application db over BF scheme. Application db is a simple program with only three classes and seven reference fields. The reference fields most accessed by the benchmark were captured by the DF scheme. It is also important to note that db puts considerable strain on the memory subsystem and has much higher cache and DTLB misses. It has a much larger working set, which thrashes the TLB. Reordering objects through DF scheme reduced L1 and L2 cache misses significantly.

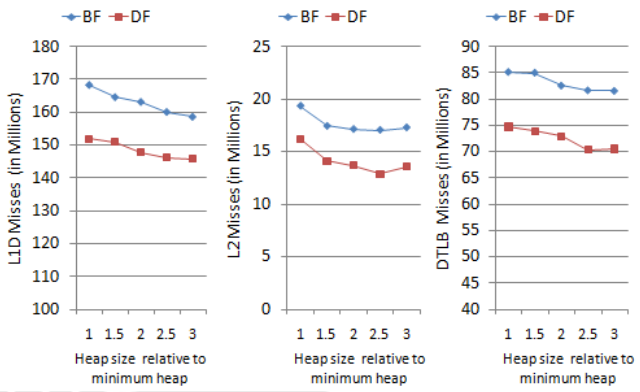


(a) Mutator

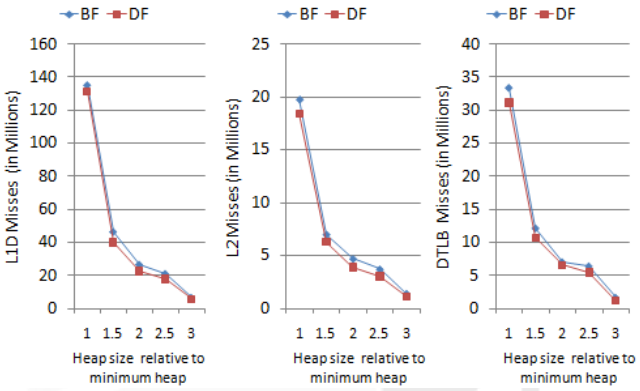


(b) Garbage collector

Figure 8. L1D cache, L2 cache and DTLB misses for db



(a) Mutator



(b) Garbage collector

Figure 9. L1D and L2 cache misses for javac

Figure 8 and 9 show the impact of BF and DF schemes on cache and DTLB misses for mutator and garbage collector. We present the results for db and javac for different heap sizes since the other applications show a similar trend to these applications. The results show a consistent effect of BF and DF schemes on mutator cache and DTLB misses over different heap sizes and for the garbage collector, the effect is noticeably low.

5. RELATED WORK

Research on garbage collector and the interplay of garbage collection with the memory subsystem are related to this work. Blackburn at al. [5] compared mark-sweep, copying, and reference counting collectors and found that for moderate heap sizes, copying garbage collector had better performance than others. They also noted that object allocation order had an advantage over segregating by size. Herts et al. [11] looked at the interaction of garbage collection with paging on real hardware. Guyer et al. [9] allocated connected objects in the same garbage collection memory space. They performed static compile-time analysis to determine in which memory space the source object resides, and then allocates the target object to the same memory space. Shuf et al. [15] studied the interplay of garbage collection with the memory subsystem, without specifically looking at different object orderings.

6. CONCLUSIONS

Semi-space garbage collector provides opportunity for object ordering evaluation through its garbage reclamation. We implement two pre-determined object ordering schemes to evaluate the impact of different object ordering in memory on the performance of Java applications. Our experiment results show that the ordering of objects in memory has an impact on Java application execution time, cache memory and DTLB misses. Object ordering in memory might improve or degrade the performance of application, cache and DTLB. As expected, the order of objects in memory affects the number of cache and DTLB misses more than the execution time of the applications.

7. REFERENCES

[1] Alpern, B., Attanasio, C. R., Barton, J. J., Burke, M. G., Cheng, P., Choi, J.-D., Cocchi, A., Fink, S. J., Grove, D., Hind, M., Hummel, S. F., Lieber, D., Litvinov, V., Mergen, M. F., Ngo, T., Russell, J. R., Sarkar, V., Serrano, M. J., Shepherd, J. C., Smith, S. E., Sreedhar, V. C., Srinivasan, H., and Whaley, J. 2000. The Jalape˜no Virtual Machine. IBM System Journal 39, 1, 211–238.

[2] Alpern, B., Augart, S., Blackburn, S., Butrico, M., Cocchi, A., Cheng, P., Dolby, J., Fink, S., Grove, D., Hind, M., McKinley,

K., Mergen, M., Moss, J., Ngo, T., Sarkar, V., and Trapp, M. 2005. The Jikes Research Virtual Machine Project: Buliding an Open-source Research Community. IBM Systems Journal 44, 2, 399–417.

[3] Abuaiadh, D., Ossia, Y., Petrank, E., and Silbershtein, U. 2004. An efficient parallel heap compaction algorithm. In OOPSLA `04.

[4] Blackburn, S. M., et al. 2006. The DaCapo benchmarks: Java benchmarking development and analysis. In OOPSLA`06.

[5] Blackburn, S. M., Cheng, P., and McKinley, K.S. 2004. Myths and realities: The performance impact of garbage collection. In SIGMETRICS, 32,1,25-36.

[6] Collins, G. E. 1960. A method for overlapping and erasure of lists. Commun. ACM, 3, 12, 655-657.

[7] Chen, Y., Dios, R., Mili, A., Wu, L., and Wang, K. 2005. An empirical study of programming language trends. IEEE Software, 22, 3, 72-78.

[8] Eeckhout, L., Georges, A., and De Bosschere, K. 2003. How Java programs interact with virtual machines at the microarchitectural level. In OOPSLA `03, 169-186.

[9] Guyer, S. Z. and McKinley, K. S. 2004. Finding Your Cronies: Static Analysis for Dynamic Object Colocation. In OOPSLA ’04, 25–36.

[10] Hertz, M., Blackburn, S. M., Moss, J. E. B., McKinley, K. S., and Stefanovi´c, D. 2002. Error-free Garbage Collection Traces: How to Cheat and Not Get Caught. In SIGMETRICS ’02, 140–151.

[11] Hertz, M., Feng, Y., and Berger, E. D. 2005. Garbage collection without paging. In PLDI `05.

[12] Jones, R. E., Lins, R. D. 1996. Garbage collection: algorithms for automatic dynamic memory management. Wiley, Chichester.

[13] McCarthy, J. 1981. History of LISP. In ACM History of programming language 1, 173-185.

[14] Standard Performance Evaluation Corporation. http://www.spec.org/benchmarks.html#java

[15] Shuf, Y., Gupta, M., Franke, H., Appeal, A., and Singh, J.P. 2001. Characterizing the memory behavior of Java workloads: A structure view and opportunities for optimizations. In SIGMETRICS ’01.

Using Data Mining to Improve Prediction of 'No Show' Passenger on An Airline Reservation System

Johan Setiawan
Universitas Multimedia Nusantara
Scientia Garden, Boulevard Street
Gading Serpong, Tangerang, Banten
+62.817.678.0889
johansetiawan@unimedia.ac.id

Bobby Limantara
Universitas Bina Nusantara
Jl. KH Syahdan No. 9
Kemanggisan, Palmerah, Jakarta
+62.21.535.0660
bobbylimantara@yahoo.com

ABSTRACT

One of the problems facing the airline industry is to predict number of passengers will go on the departure time but somehow they do *'no show*'. This is known as '*No-Show'* Passengers. Accuracy in predicting number of *'no show'* passengers will increase airlines profit because an *empty seat* prediction can be lowered, *no show* and *denied boarding* causes by over prediction number of passenger *'no show'* can be avoided

The purpose of this research is to *design a predictive model* using data mining at PT Metro Batavia to predict 'no show' passenger.

Methodologies used in this research are: *analyzing current business process and model*, *design model*, *implementation and evaluation model*. In designing the predictive model, specific information about PNR (*Passenger Name Record*) becomes the *input* for the model. Oracle Data Miner is used as an *implementation model* using data mining *classification* and *Naïve-Bayes* algorithm. The Evaluation model use *mean absolute errors*. Based on the evaluation, predictive model built has a lower error rate compare with current prediction model used at PT Batavia Air. In conclusion, the implementation of predictive model airline *no show rate* based on PNR can improve accuracy in predicting *'no show'* passenger at PT Metro Batavia

Keywords

Data Mining, Predictive Model, Classification, Naive Bayes, Airline, No show rate, Passenger Name Record

1. INTRODUCTION

Information technology development has brought a lot of changes for the human kind, including the business world. Currently, information technology has already become a requisite for big and small companies to stay competitive.

One of the technologies most business use to help the business process is database which can help to record daily transaction. Unfortunately many times data only accumulated in the database and has not much used, then data becomes "data tombs". Many companies have a lot of data and it contains giant information inside it. It becomes so much and makes more difficult to dig the information by using *traditional analysis* method.

One of the many way to analyze the hiding information and many times can not be seen by eyes is using *data mining techniques*. Data Mining can be use to dig hidden information inside the data. It can found hidden pattern in big and complex data, pattern that usually cannot be solve with *analysis approach* and *traditional statistic* cause of many attributes or the pattern too complex.

PT Metro Batavia is an airlines company using **Batavia Air** brand name. Like many other airlines, Batavia Air also implement *overbooking system* -- a system which allow passengers to book seat capacity on a flight that is more than the capacity of the flight.

Some airlines routinely do the overbooking mechanism in an expectation there are some passengers already booked the seat but *no show*. Accuracy in predicting number of no show passengers can increase the profit by reducing the spoiled seat -- a seat that has to be sold, and reducing the number of involuntary denied-boarding.

Current *conventional method* use by Batavia Air, to predict number of *no show* passengers is using ***average no show rate*** based on the historical data on the same flight, unfortunately the use of the method is not quite accurate.

The research written here is based on the necessity way to improve the accuracy on predicting *no show* passengers, to create a *predictive model* using data mining that include the specific information of every passenger that store in the *Passenger Name Record* (PNR). With the predictive model Batavia Air hope they can improve the accuracy on predicting *no show rate* and increase the profit.

2. THEORY

2.1 Data Mining

According to Connolly and Begg (2005,p1233)[1] *data mining* is a process to extract valid information, unknown before, can be understood, and actionable from big database so it can be used to make a crucial decision.

According to Han and Kamber (2006,p7)[2] *data mining* is the process of discovering interesting knowledge from large amounts of data stored either in databases, data warehouses, or other information repositories.

Based on this view, the architecture of a typical data mining system may have the following major components (Han and Kamber, 2006, pp7-8) [2]:

1. *Database, data warehouse, data mart, World Wide Web, or other repository*: This is one or a set of databases, data warehouses, spreadsheets, or other kind of repositories. Data cleaning or data integration technique maybe needed to prepare the data.

2. *Database or data warehouse server:* Database or data warehouse server is responsible for fetching the relevant data based on the user's data mining request.

3. *Knowledge base:* This is the domain knowledge that is used to guide the search or evaluate the interestingness of resulting mining. Such knowledge can include *concept hierarchies* used to organize attributes or attributes value into different level abstraction. Knowledge such as user beliefs, which can be used to assess a pattern's interestingness based on its unexpectedness, may also be included. Other examples of domain knowledge are additional interestingness constraints, or thresholds, and metadata.

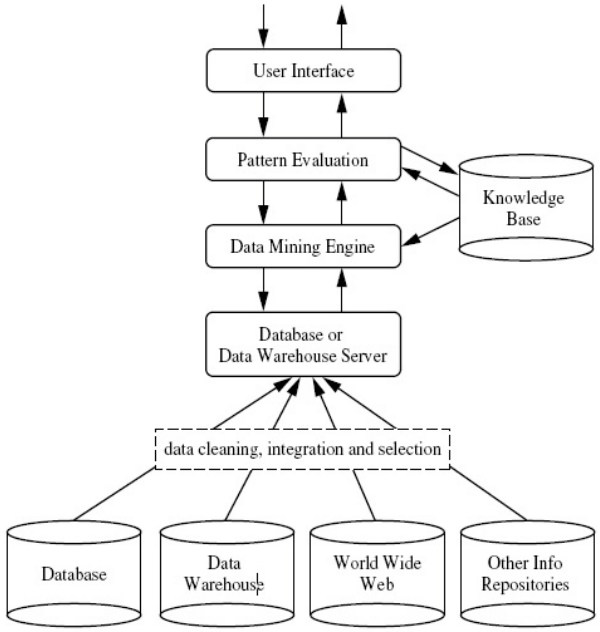


Figure 1. Data Mining major components

4. *Data mining engine:* The essential part of the data mining system and ideally consist of a set of functional module for tasks such as characterization, association analysis, classification, evolution and deviation analysis.

5. *Pattern evaluation module:* This component typically employs interestingness and interacts with the data mining modules so as to focus the search towards interestingness patterns. It may access interestingness thresholds stored in the knowledge base. Alternatively the pattern evaluation module may be integrated with the mining module, depending on the implementation of the data mining method used.

6. *User interface:* This module communicates between users and the data mining system, allowing the user to interact with the system by specifying a data mining query or tasks. This component allow user to search database and data warehouse or data structure, evaluate mining pattern, and visualize the patterns in different format.

2.2 Classification

Classification is a form of data analysis used to create a model describing data class to predict class for new data.

Classification predicts categorical value in example value in no order, and discrete based on the vector attribute. Algorithm that can be used for classification such as: Naïve Bayes, decision tree and support vector machine.

Classification consists of 2 (two) process (Han and Kamber, 2006, pp285-288) [2]: *learning phase* and *classification*. The first step, *classifier* (predictive model which predict categorical class value) created to describe data class previously defined. This learning phase is a phase where classification algorithm create predictive model by learning training set that consists of database record and class label.

A record X, represented by n-dimension vector attributes, X= (, ,…,) where , … is attribute value , ,..., . Every record, X, assume join inside a class previously defined through other database attribute, class label attribute. Class label attribute is a discrete value and has no order. Class label attribute value is a categorical where every possible function as category or class.

Because every class labels at each training records already known, this phase is also called *supervised learning*. The purpose of supervised is the learning process of classifier watched, supervised which classifier given to class where training record join. This is contrary with *unsupervised learning* where label class not known, and number of classes to be learned previously unknown.

The first step of classification process can be named as *learning function*, *y=f(X)*, which can predict class label *y* if record given. Classification tries to learn function or mapping to separate data class.

The second phase of classification process is to *test the model* where model use as *classification*. The purpose of second step is to *measure accuracy of classifier*. Data input for the test should not be using the same data as *the training set*. The classifier test result using the same data training is not a good indicator for the classifier performance. This is because the classifier created using the same data at the test time so the estimation performance result is *optimistic*. Error rate from evaluation result from *training data* called *resubstitution error*. Classifiers tend to over fit data because at *the learning phase* classifier may include some anomalies at training data that is not at the overall general data. Because of that, the *test set* that is used produced from different records *training set* where record not use to create *classifier* (Witten and Frank, 2005, p.145)[3]

2.3. Bayesian Classification

Bayesian *classifier* is a *classifier statistic* which can predict member probabilities of a class, such as probability one record join in certain class.

Bayesian Classification based on Bayes theorem has a high level accuracy and can be run fast at a big database.

Naïve bayesian classifier assumed effect of an attribute value at one class, *independent* from other attribute value. This assumption is also called *class conditional independence*. This assumption use to simplify computational process and therefore called '*naïve*'.

If *X* is a data record, where *X* consists of *n* attributes in bayesian terminology, *X* named as *facts*. If *H* is a hypothesis, for example record *X* is a member of class C.

For classification, determined *P(H|X)*, probability hypothesis *H* if facts given or record *X*.

In other words, the search is probability record *X* a member of class *C*, given attribute description of *X*. *P(H|X)* is a *posterior probability*, *H* conditioning at *X*.

As an example, if customer data in one computer store describe by age and income attribute. *X* is a customer, age 35 years and has an income of Rp. 10.000.000 then if we want to know either *X* will buy the computer given the age and income. *P(H)* is *prior probability*.

In the above example, probability customer wants to buy computer without looking at the age and income or other information. *P(X|H)* is *posterior probability* where *X* conditionalizes to *H*. As per example, the probability *customer X*, age 35 with income Rp. 10.000.000 if customer bought the computer.

P(X) is a *prior probability* from *X*. Using the example, it means *probability* someone from customer database, with age 35 and income Rp. 10.000.000,00.

P(H), *P(X|H)*, and *P(X)* can be found from *training set* where train set that has class label. Bayes theorem is useful to count *posterior probability P(H|X)* from *P(H)*, *P(X|H)*, and *P(X)* with this formula

$$P(H|X)=\frac{P(X|H)P(H)}{P(X)}$$

Bayesian Classifier work process as follow:

1. If *D* is a *training set* consists of record and label of their classes. Every record represented with *n*-dimension *attribute vector*, *X*= (, ,…,) and has *n* attribute , ,…, .

2. If there are *m* class, , ,…, . Given the record X, *classifier* will predict *X* as a member of class that has highest *posterior probability* value, conditioning at *X*. Naïve Bayes *classifier* predict record *X* joined in class if and only if

$$P(\ |X)>P(\ |X) \quad for\ 1\leq j\leq m, j\neq i$$

Then *P(|X)* value is the *highest probability value*. The value of class where *P(|X)* maximized called as *maximum posteriori hypothesis*.

$$P(C_i|X)=\frac{P(X|C_i)P(C_i)}{P(X)}$$

3. Because *P(X)* value constant for all classes, then only $P(X|C_i)I$ that should be maximized. If *prior probability* of class unknown, then usually assumed that every classes is the same which *P( )=)=…= P*, and the one should be maximized only $P(X|C_i)$ value. Besides that, $P(X|C_i)I$ value should be maximized. *Prior probability* class value can be estimated with $I=|\ |/|D|$, where | | is number of record inside *D* that has class label .

4. When dataset given have a lot of attributes, then it will be very difficult and costly to calculate $P(X|C_i)$ value. To reduce computational process in evaluating $P(X|C_i)$, *naïve class conditional independence* assumption created. The assumption treats value of an attribute *independent* from one to another. Then

$$P(X|C_i) = \prod_{k=1}^{n} P(X_k|C_i)$$

$$= P(X_1|C_i) \times P(X_2|C_i) \times \dots \times P(X_n|C_i)$$

Probability of $P(X_1|C_i)$, $P(X_2|C_i)$,…, $P(X_n|C_i)$ can be search from training data. show attributes value for record *X*.

For each attribute, will be seen if the attributes is *categorical* or *continues* value. For example, to calculate *P(X|* , do as follows:

If is *categorical*, then $P(X_k|C_i)$ is number of record that has label class in *D* and has value for attribute divided by | | which is the number of record with class label pada *D*

If *is continue* value, then an additional calculation needed.

Attribute with *continue* value will be assumed has *Gaussian distribution* with mean dan standard-deviation , which is defined:

$g(x, \ , \ ) = \frac{1}{\sqrt{2\pi\sigma}}e^{-}$ then

P(|) = g(, ,)

We need to calculate (mean) dan (standard deviation) from attribute value for record with class.

For example, if *X* = (35, Rp. 10.000.000), where age attribute and income attribute. Class label is *buys_computer*. Label class value for *X* is *yes*. If age attribute is not *discretisized* and fixed as attribute with *continue* value. Suppose from *training set*, founded customer inside *D* who bought computer has an age between $38^{\pm}12$. Then in other words for general attribution this class has value of μ : and =*12*. The value of μ : and =12 used to estimate *P* (age=35|buys_computer=yes).

5. To predict class label *X* then $P(X|C_i)P(\ )$ evaluated for each class . Classifier predicted class label from record *X* is class if and only of

$$P(X \ ) P(\ > P(X \ ) P(\quad for\ 1\leq j\leq m, j\neq i$$

In other words, prediction class value for record *X* is classes where *P(X)P(* is maximized. (Han and Kamber, 2006, pp310-313)

3. ANALYSIS OF CURRENT MODEL

3.1 Seat Overbooking Allocation

Number of seat allocated for all class flight usually more than physical capacity (*overbooking*), because usually there are passengers *cancel* or *no show*. Total seat capacity allocated calculated as flight cabin capacity plus number of passenger 'no-show' estimated for that flight.

RC staff predicts number of passengers *'no-show'* for a flight using *historical model* and intuition. Historical model predict *'no-show'* passenger at a future flight calculated from the *average of no show* passengers from the collection of historical data flight for the same route

$$\rho_{hist}^{m}=\frac{\sum_{k\in N_m} NS^k}{\sum_{k\in N_m} T^k} \text{ where}$$

m shown the unique flight, characterize by flight number, route, departure date and ETD (estimated time departure)

k shown the unique flight at historical data, characterize by flight number, route, departure date and unique ETD

prediction of *no show* rate from historical model (hist) at flight *m*

number of *no show* at flight *k*

number of passengers already buy the ticket (already issued the ticket) at flight *k*

shown a collection of historical data flight which has the same flight route at flight *m* on certain period

In the formula above, *no show rate* at flight *m*, ρ_{hist}^{m} predicted by calculating number of no show passengers at historical collection of no show passenger , N_m , which has the same route at flight *m* during certain period, divided by number of passengers ticketing at historical collection flight, N_m which has the same route as flight *m* during certain period.

Using historical model, prediction of number of *no show* passenger at flight *m*. NS_{hist}^{m} , can be obtained from:

$$NS_{hist}^{m} = \rho_{hist}^{m} T^{m}$$

Where: prediction *no show rate* at flight *m* based on historical model and number of passengers ticketing at flight *m*.

3.2 Weaknesses of Historical model

Prediction model *no show* currently implemented at PT Metro Batavia is *historical model*. Here is the formula

$$\rho_{hist}^{m} = \frac{\sum_{k \in N_m} NS^{k}}{\sum_{k \in N_m} T^{k}}$$

If an overbooking wants to be implemented at a flight *m* then *no show rate* determine from number of *no show* passenger at N_m and divided by number of *ticketing* passenger at N_m.

Where N_m is a collection of historical data flight

From the above historical model, shown that factor(s) or attribute included in determining *'no-show' rate* only from flight route. Other factors like flight day, departure time, class flight bought by passengers are not included in the calculation.

Table 1 shown number of 'No Show' and 'Ticketing' during the period of January-July 2008 based on the route can be seen below

Using the above *historical model* formula, then if we want to calculate *overbooking* for flight *m* route *CGK-MES* will always get no show rate **0.035070913** for the same period

Table 1. No Show rate based on route

Route	No Show	Ticketing	No Show Rate
CGK-MES	**3185**	**90816**	**0.035070913**
MES-CGK	**2283**	**89223**	**0.025587573**
CGK-SUB	**3707**	**74770**	**0.049578708**
PNK-CGK	**1536**	**74405**	**0.020643774**
SUB-CGK	**3134**	**64721**	**0.048423232**
PKU-CGK	**1327**	**50513**	**0.026270465**
CGK-PKU	**1865**	**49022**	**0.038044143**
CGK-BTH	**2340**	**51565**	**0.045379618**
BTH-CGK	**1203**	**44631**	**0.026954359**

Weaknesses using *historical model* is lack of factors or attributes included in the calculation of *no show rate*. If there is a *trend* or certain pattern affect the *no show rate* on another attributes such as day of departure, or flight class then the prediction result becomes not accurate. Table 2 show the *No Show Rate* based on day of departure for each route:

Table 2. No Show Rate based on day of departure

Route	Senin	Selasa	Rabu	Kamis	Jumat	Sabtu	Minggu
CGK-MES	0.02262	0.03986	0.05403	0.03742	0.04072	0.03178	0.01881
MES-CGK	0.01445	0.02745	0.04079	0.02392	0.03561	0.02486	0.01156
CGK-SUB	0.03833	0.05227	0.07552	0.04496	0.05204	0.04781	0.03422
PNK-CGK	0.00635	0.02967	0.03805	0.02106	0.02889	0.01383	0.00619
SUB-CGK	0.02642	0.05663	0.08481	0.04245	0.05061	0.05154	0.03073
PKU-CGK	0.01406	0.03127	0.03995	0.0308	0.03207	0.02138	0.0133
CGK-PKU	0.02884	0.04638	0.05788	0.03645	0.03807	0.03458	0.02417
CGK-BTH	0.03262	0.06321	0.05915	0.03471	0.05554	0.03952	0.03264
BTH-CGK	0.0128	0.03675	0.045	0.02909	0.038	0.01882	0.009
[illegible]	0.021832	0.04261	0.05502	0.033429	0.041283	0.031569	0.020069

Based on the table above there is a tendencies on Tuesday, Wednesday and Friday have *no show rate* higher than on Sunday and Monday. For example route CGK-MES if *the historical model* use either it is on Sunday or Wednesday, the prediction *no show rate* is 0.035070913.

This will become a problem, *overprediction* on lower *no show rate* on Sunday and Monday, it will cause *underprediction* for high *no show rate* on Tuesday, Wednesday, Thursday and Friday.

Table 3 below shown that for a certain classes have a *no show rate* higher compare with other classes. Class P, D, H, N, W, Y, and Z has a higher *no show rate* (NSR>4%) compare with class B, L, M, T, and V (NSR<2%).

Using *historical mode*l, if there are many passengers booked for classes P, D, H, N, and W then *underprediction* can be used because the *historical model* does not include the class information. On the contrary, if there are many passengers booked for classes B, L, M, T, and V then *overprediciton* happened.

Table 3 below will show *no show rate* based on 16 flight classes currently available at PT Metro Batavia:

Table 3. No Show rate based on 16 flight classes

KELAS	CGK-MES	MES-CGK	CGK-SUB	PNK-CGK	SUB-CGK	PKU-CGK	CGK-PKU	CGK-BTH	BTH-CGK	[illegible]
B	0.03739	0.02439	0.02649	0.01571	0.03854	0.02422	0.02412	0.02685	0.01992	0.026403
D	0.06425	0.03224	0.02941	0.02763	0.07377	0.03797	0.06015	0.0354	0.02362	0.042716
H	0.03016	0.05027	0.01902	0.04492	0.09192	0.02844	0.04502	0.0674	0.02874	0.045099
L	0.02025	0.02628	0.0212	0.024	0.02527	0.01542	0.00728	0.01894	0.01571	0.019372
M	0.0378	0.0369	0.05068	0.03018	0.03166	0.02547	0.0356	0.03118	0.01844	0.033101

N	0.04578	0.0251	0.05341	0.04355	0.04852	0.02597	0.0546	0.06318	0.03046	0.043397
P	0.06589	0.03808	0.08324	0.03578	0.07818	0.04118	0.05659	0.06991	0.04165	0.056722
Q	0.03751	0.02182	0.043	0.00711	0.03893	0.01773	0.02114	0.03314	0.02139	0.026863
R	0.03099	0.01884	0.03057	0.01059	0.03325	0.01704	0.0287	0.04075	0.02081	0.025727
S	0.02488	0.02133	0.03417	0.02084	0.02824	0.01463	0.02792	0.02526	0.02343	0.024522
T	0.0121	0.01484	0.03077	0.00612	0.02865	0.01453	0.02065	0.0388	0.01607	0.020192
V	0.01986	0.01603	0.03114	0.01608	0.02875	0.01676	0.02823	0.03922	0.01681	0.023653
W	0.06876	0.04876	0.07782	0.02497	0.07189	0.04591	0.06276	0.07365	0.03806	0.056953
X	0.02426	0.0174	0.03226	0.02097	0.04487	0.02304	0.04014	0.05375	0.03078	0.031941
Y	0.07138	0.04924	0.11685	0.01408	0.1073	0.03413	0.03882	0.02762	0.02628	0.053967
Z	0.02041	0	0.11475	0.08333	0.07684	0.0282	0.06823	0.08716	0.03078	0.056633

4. PROPOSED PREDICTIVE MODEL

Predictive model *no show rate* based on PNR will be created using Data mining function *classification*. Classification use because the purpose of this model is to *predict* a class label *value* and class label *attribute values* in categorical (*show* or *no show*)

The first step is to define vector attribute shown the characteristics of every passengers in a flight. *Capital alphabet* will be use to shown the vector attribute and *lowercase alphabet* for the value. If X_i. i=1.... I shown I *vector attribute* for each passengers. Then combining all the *vector attributes* will gave a vector.

$$X=[X_i \ldots X_I]$$

For every passengers, n=1, ... I which already ticketing at flight *m*, represented with vector from *vector attribute*

$$x_n^m=[x_{n,1}^m, x_{n,2}^m,\ldots, x_{n,I}^m]$$

A class label *C* with value c_n^m shown if the passenger *show* or *no show* (NS). The predictive model work as follow: if given a set of class label c_n^m and a set of *vector attribute* vector x_n^m model will predict the output class probabilities from passengers *n* at flight *m*

$$P(C=c_n^m|X=x_n^m)$$

Because we want to predict the probability of *no show*, c_n^m=*NS*, then *no show* probability for passenger *n* at flight *m* can be written as

$$P(NS|x_n^m)$$

P(NS|x_n^m) calculation using Naive Bayes. The reason to choose Naive Bayes because this method allows to uses a very big data set and the calculation is very fast.

Using Naive Bayes, then

$$P(NS|x_n^m)=\frac{P(x_n^m|NS)P(NS)}{P(x_n^m)}$$

Because Naive Bayes using an assumption that naive class *conditional independence* then the probability of *no show* a passenger can be written as

$$P(NS|x_n^m)=\frac{\prod_{k=1}^{I}P(x_{n,k}^m|NS)P(NS)}{\prod_{k=1}^{I}P(x_{n,k}^m)}$$

$$P(NS|x_n^m)=\frac{P(x_{n,1}^m|NS)\times P(x_{n,2}^m|NS)\times \ldots \times P(x_{n,I}^m|NS)\times P(NS)}{P(x_{n,1}^m)\times P(x_{n,2}^m)\times \ldots \times P(x_{n,I}^m)}$$

Predictive model built to predict number of *no show* passengers at one flight with

$$NS_{model}^m=\sum_{nem}P(NS|x_n^m)$$

At the above formula, *No show* passengers at flight *m* can be calculated from the sum of *no show* passenger probability for ticketing at flight *m*.

If T_m shown the number of passengers ticketing at flight *m* then *No show* rate, ρ_{model}^m model, at one flight can be shown using these equation

$$\rho_{model}^m=\frac{NS_{model}^m}{T_m}$$

4.1 Input Data Model

Table 4. Data Input Model Design

Attribute Name	Data type	Attribute type	Value
ID	Varchar	-	-
No_ show	number	categorical	2
Kota_berangkat	varchar2	categorical	6
Kota_tujuan	varchar2	categorical	6
Kelas	varchar2	categorical	16
Segmen	number	categorical	2
Seat	number	numerical	20
Gender	number	categorical	3
Day_derpature	varchar2	categorical	7
Time_of_Day	varchar2	categorical	6 (binned)
PNR_split	number	categorical	2
PNR_rebooked	number	categorical	2

ID is a *unique id* for a record, *No Show* field is a *target attribute*, other fields are *vector attributes.*

All *vector attributes* and *target class* will be extracted for each passengers, for example if in one PNR (Passenger Name Record) there are 4 passengers then it will shown 4 (four) record for input to data mining model.

5. IMPLEMENTATION AND EVALUATION

5.1 Implementation

The implementation will be using Oracle Data Miner with *classification* function. Before starting the design process model, *relevance analysis* was done with attribute *subset selection*. The purpose of doing the attribute subset selection is to identify if any attribute vector that has no contribution for determining the *target attribute* on *classification* process. Attribute that has no influence to the *target attribute* will be eliminated from *classification* to make the mining process faster.

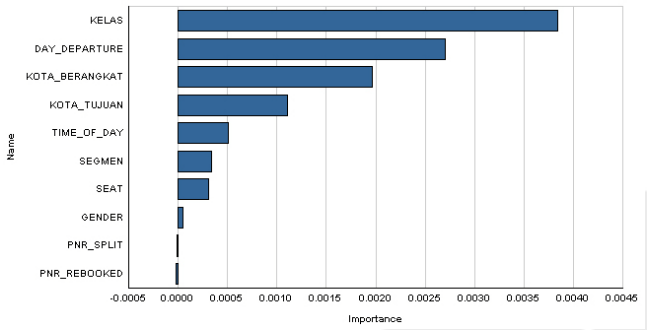


Figure 2. Attribute Importance

The result of *attribute importance* above shown attribute *class* has significances to the *target attribute*. Besides that, *day of flight* and *departure city* also has a big influence to the target attribute.

5.2 EVALUATION for Predictive Model

Evaluation for *predictive model* is done by calculating the error from the model. Measurement instrument use *Mean Absolute Error*, because *Mean Absolute Error* produce value with the same original value and not increasing the *outlier*

Error *No Show Rate* calculation from the model calculated with

$$\epsilon_{model}(\rho)=\frac{1}{N_f}\sum_{m=1}^{N_f}|\rho_{model}^{m}-\rho_{aktual}^{m}|$$

where

$\epsilon_{model}(\rho)$ *mean absolute error no show rate* value for model

N_f number of *flight* at *evaluation set*

ρ_{model}^{m} prediction *no show rate* model at *flight m*

ρ_{aktual}^{m} *no show rate* at the real data

Error calculation no show passengers from model calculated using $\epsilon_{model}(NS)=\frac{1}{N_f}\sum_{m=1}^{N_f}|NS_{model}^{m}-NS_{aktual}^{m}|$ *where*

$\epsilon_{model}(NS)$ *mean absolute error* value for passengers *no show* for model

N_f number of *flight* at *evaluation set*

NS_{model}^{m} prediction number of *no show* passengers from model at *flight m*

NS_{aktual}^{m} number of *no show* actual passengers at *flight m*

By using *historical model* as a basis of calculation, then accuracy improvement on *no show* prediction, δ_{model}(NS), for *predictive model* compare with current historical model

$$\delta_{model}(\text{NS}) = \epsilon_{hist}(NS) - \epsilon_{model}(NS)$$

Table 6 Evaluation model summary

Model	$\epsilon_{model}(\rho)$	$\epsilon_{model}(NS)$	δ_{model}(NS)
Historical	**0,0329894**	**3,63565**	-
PNR Based (decision tree)	**0,0283048**	**3,16382**	**0,47237**
PNR Based (naïve bayes)	**0,0253441**	**2,81057**	**0,82508**
Number of *Flight*: 2228			

$\bar{\rho}_{actual}$ = **0,025296**

$\overline{NS}_{actual}$ = **3,05116696**

Table 6 shown the accuracy of *predictive model* using specific information about PNR (*Passenger Name Record*) can predict number of *no show* better than the *historical model* currently used. Besides using Naive Bayes as a *comparison* for the test, *decision tree algorithm* use. The performance result has shown *Naive Bayes* better than *decision tree* algorithm. Improvement accuracy *predictive model* based on PNR (*Passenger Name Record*) compare with the current system in average *0.82508* passengers on every flight or around *one passenger* for every flight.

5.3 System Design

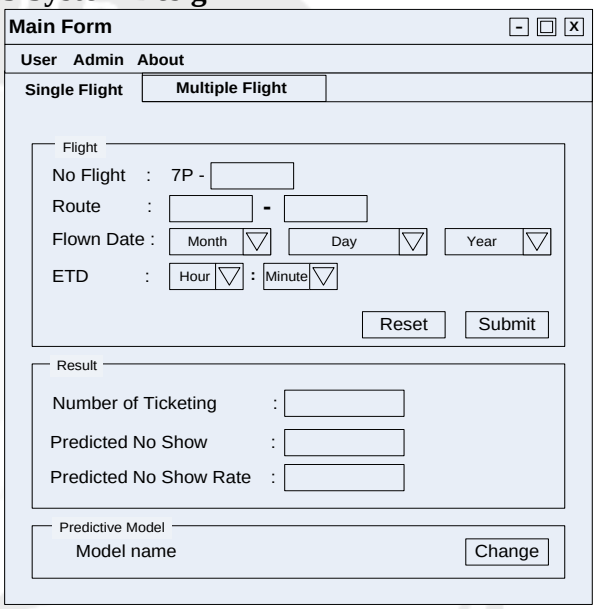


Figure 3 Main Form for Single Flight

Form design on Figure 3 is the main form the user saw when the user successfully *login*. In this form user can type information about a flight and after clicking on the submit button, the result will be shown below the input.

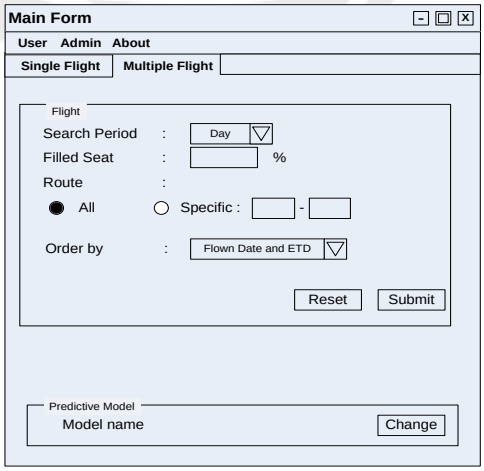


Figure 4 Main Form for multiple flights

Figure 4 is the Main Form for multiple flights. In this form the user can search for multiple flights at the same time based on the

input type. After clicking the submit button, the result will be show in Figure 5 below.

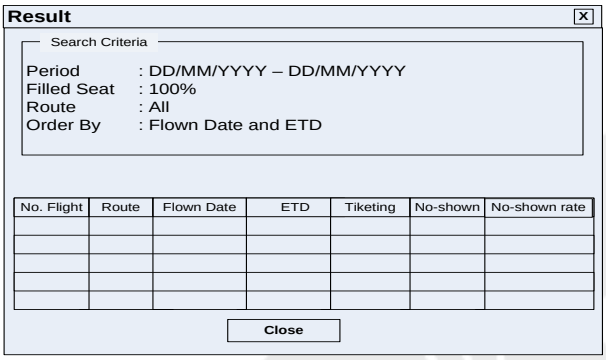


Figure 5 Result for multiple flights

6. CONCLUSION

Based on the analysis and design result, the conclusion:

1. Predictive Model Airline *No show rate* based on Passenger Name Record built can predict *No show* Passenger more accurate than the *historical model* on the current system

2. The use of specific passenger information on the Passenger Name Record *can improve* the accuracy of predicting *no show* passenger.

3. User Interface application built can ease Reservation Control (RC) staff to use *predictive model* built.

7. REFERENCES

[1] Connoly, Thomas, Begg, Carolyn (2005). *Database System: A Practical Approach to Design, Implementation, and Management*, Fourth edition. Addison Wesley, Essex.

[2] Han, Jiawei, and Micheline Kamber.(2006).*Data Mining Concept and Techniques*, second edition. Elsevie, San Fransisco.

[3] Witten, Ian H, and Frank, Eibe (2005). *Data Mining: Practical Machine Learning Tools and Techniques*, Second edition, Morgan Kaufmann Series in Data Management Systems.

[4] Anonym. (2008).*Business Intelligence Introduction*. http://www.learndatamodeling.com

[5] Anonym. (2008). http://en.wikipedia.org/wiki/Airline

[6] Anonym. (2008). http://en.wikipedia.org/wiki/Computer_reservations_system

[7] Anonym. (2008). http://en.wikipedia.org/wiki/Overbooking

[8] Anonym. (2008). http://en.wikipedia.org/wiki/PNR

[9] Anonym. (2008). http://en.wikipedia.org/wiki/Record_locator

[10] Berson, Alex, and Stephen J. Smith. (1997). *Data Warehousing, Data Mining, & OLAP*. McGraw-Hill, New York. U.S.A.

[11] Chapman, Pete et al.(2000).*CRISP-DM 1.0 Step-by-step Data Mining Guide*. CRISP-DM Consortium.

[12] Inmon, W. H. (2002).*Building The Data Warehouse*, 3rd edition, Wiley. Computer Publishing, USA.

[13] Laudon, K and Laudon, J (2000). *Management Information System: New Approach to Organization and Technology*, Fifth Edition. Prentice Hall, New Jersey.

[14] Oracle, 2005, *Oracle Data Mining Concept*, Oracle.com.

Using Frequent Max Substring Technique for Thai Keyword Extraction Used in Thai Text Mining

Todsanai Chumwatana
School of Information Technology
Murdoch University
South St, Murdoch
Western Australia 6150
T.Chumwatana@murdoch.edu.au

Kok Wai Wong
School of Information Technology
Murdoch University
South St, Murdoch
Western Australia 6150
K.Wong@murdoch.edu.au

Hong Xie
School of Information Technology
Murdoch University
South St, Murdoch
Western Australia 6150
H.Xie@murdoch.edu.au

ABSTRACT

The amount of electronically stored information in Thai language has grown rapidly in the past few years and the number of these documents is still increasing. This makes information extraction (IE) an essential task for extracting keywords from Thai texts. Thai texts are considered as un-delimited language where the structure of writing is a string of symbols without explicit word delimiters. Words in Thai language are not naturally separated by any word delimiting symbols. Due to this characteristic of Thai written language, word segmentation is a challenging task and has become one of the important research topics. Many word segmentation techniques have been proposed to segment Thai texts into a set of words to support extraction of keywords. However, most of the word segmentation approaches required complex language analysis. They usually rely on language analysis or on the use of dictionary or corpus. In this paper, an alternative method for extracting important Thai keywords is proposed. The proposed approach is based on the analysis of frequent max substring that extracts important keywords. This approach looks for long and frequent substrings rather than individual words from given texts. As a result, this approach is language-independent. It does not rely on the use of dictionary or language analysis. We refer this technique as Frequent Max substring mining or FM technique. Applying the FM technique to Thai texts yields a set of keywords that are frequent and highly distinct from given texts. The set of extracted keywords from FM technique is able to contain all frequent substrings without information loss. Therefore this technique uses less space for storing all frequent substrings in order to support the growth of Thai electronic information.

Keywords

Frequent Max Substring Mining, Text Mining, Information Extraction

1. INTRODUCTION

Text mining, sometimes referred to as text data mining, is the application of data mining for text processing and Information Retrieval (IR). Text mining refers to a process of deriving useful information or knowledge from texts [1], [2]. Information Extraction (IE) is one essential task in the area of text mining that describes a process of discovering interesting keywords underlying unstructured natural-language texts. Majority of the proposed methods in the literature for extracting keywords were accomplished by constructing a set of words from given texts. Keywords will then be selected from the set of words during the preprocessing step. This makes these methods work well with European languages where texts are naturally segmented into individual words by word delimiter such as white space or other special characters. However, these algorithms cannot be directly applied for Thai language. Unlike European languages, Thai language is considered as a non-segmented language where words are a string of symbols without explicit word boundaries, and also the structure of written Thai language is highly ambiguous. Thai sentence is consisting of several words using a string of characters without word separators such as white spaces, and in some cases semicolons and commas to separate these words. Due to this problem, IE is one of the essential techniques that is applied for extracting keywords from Thai texts [3], [4], [5]. Most Thai IE techniques are based on word segmentation that is one of the most widely used information extraction techniques in Natural Language Processing (NLP). The exploitation of word segmentation techniques for IE is not new and there are several approaches to Thai word segmentation as we will describe in next section. However, most word segmentation techniques usually rely on dictionary or corpus or linguistic knowledge of the language. Beside this, there are some other techniques which do not rely on language analysis such as n-grams, frequent patterns mining, and longest common prefixes techniques. These techniques do not rely on the use of dictionary or corpus and do not depend on language analysis. As a result, these techniques are most widely used to tackle many Asian languages such as Chinese, Japanese, Korea and Thai which are referred as un-delimited languages. We will review some of these approaches in detail in following section.

In this paper, an alternative method to Thai keywords extraction is proposed. The proposed approach is based on the analysis of frequent max substring [6] that extracts important keywords as long frequent substrings rather than individual words from given texts. This work is based on the method of mining sequential patterns in order to generate important keywords from the given texts. Therefore this approach works in substring (series of characters) level. As a result, this approach is language-independent. It does not rely on the use of dictionary or language analysis. We refer this technique as Frequent Max substring mining or FM technique. Applying the FM technique to Thai texts

yields a set of keywords that are frequent and highly distinct from the given texts.

2. RELATED WORKS

To extract Thai keywords from a given text, word segmentation techniques are generally the essential part of an extracting technique that usually required to segment texts into a bag of meaningful words before selecting keywords from the set of words in preprocessing phase. Keywords may be derived by computing term frequency from Thai texts. Although keywords can be specified manually by experts, this process is very time consuming and labor-intensive. Therefore, word segmentation is one technique that used to support extracting Thai keywords. Previously proposed methods for Thai word segmentation can be classified into three main categories: Dictionary-based [7], [8], Rule-based [9], [10], [11] and Machine learning-based approaches [12], [13].

However, most word segmentation techniques are language-dependent. They rely on language analysis or the use of dictionary or corpus. They also work on word-level segmentation rather than phrases or sentences.

Beside word segmentation techniques, there are some other techniques which are language-independent. In [14], an efficient algorithm for discovering optimal string patterns was proposed. However, their work is based on the method of mining association rules in order to find the proximity of segmented words in given texts. In [15], Vilo proposed an algorithm that is the generalization of the *wotd* (write-only top-down) suffix trie construction algorithm to find frequent substrings of given texts. Furthermore, another efficient method is using suffix array [16], [17] to compute term and document frequency for all substrings from texts. This technique was proposed by Yamamoto and Church in 2001 [18]. The algorithm is based on suffix arrays for computing *tf* (Term frequency) and *df* (document frequency), and many functions of these quantities for all substrings in a corpus. Additional to this technique, they also presented algorithms that make use of an auxiliary array for storing LCPs (longest common prefixes). This algorithm is much faster than the obvious straightforward implementation when the text contains long repeated substrings [19].

3. KEYWORDS EXTRACTION APPROACHES

In this section, we review the details of three approaches for keyword extraction from Thai texts: Discovering frequent patterns from string, Using suffix array to compute term frequency for all substring, and Frequent Max substring mining techniques. These three approaches are language-independent. Most techniques can be used to tackle with non-segmented texts. In order to depict the process of three algorithms with Thai text, we use Thai text = "การประกอบการ" that means "to be management" and minimum frequency = 2 to be the example text in keyword extraction.

3.1 Discovering frequent patterns from string

Discovering frequent patterns from string was proposed by Vilo [15]. This algorithm is the generalization of the *wotd* (write-only top-down) suffix trie construction algorithm to find frequent substrings of the given text. Vilo's algorithm is only interested in text patterns that occur at least *K* times of texts by constructing only the subtrees of suffix trie that correspond to the frequent substrings, as show in figure 1.

Let S = "การประกอบการ" and *K* =2

String *S* = ก า ร ป ร ะ ก อ บ ก า ร $

.pos = 1 2 3 4 5 6 7 8 9 10 11 12 13

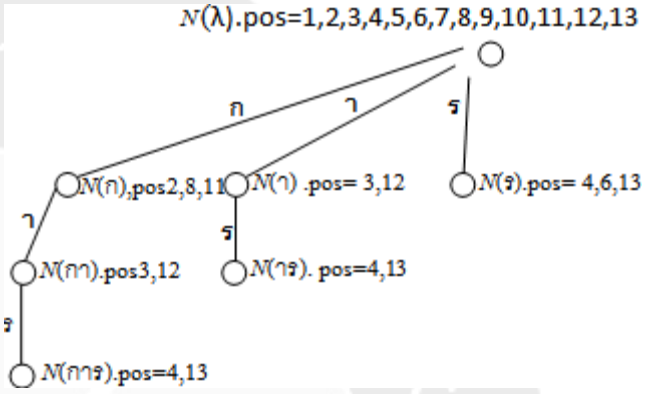


Figure 1. Discovering substring of string *S* = "การประกอบการ" having at least 2 occurrences in string S.

In Vilo's algorithm, each node in the trie represents a unique substring and contains the position list of locations in the string where substring occurs. To create the children of a node, the algorithm finds only substrings which occur at least *K* different locations of the string, only these substrings are inserted into the trie. As a result, the resulting trie contains only subtree of substrings which appear at least *K* time on different locations of texts. From figure 1, we have found that the algorithm generated six substrings from string *S* as shown in Table 1.

Table 1. The frequent substrings extracted from Vilo's algorithm

Substring	Number of occurrence
ก	3
า	2
ร	3
กา	2
าร	2
การ	2

3.2 Using suffix array

Suffix array is one efficient method to compute term and document frequency for all substrings from texts. This technique was proposed in 2001 [18] by Yamamoto and Church. The algorithm is based on suffix arrays [16] for computing *tf* (Term frequency) and *df* (document frequency), and many functions of these quantities for all substrings in texts Term frequency (*tf*) is the standard notion of frequency in corpus-based natural language

processing (NLP). It counts the number of times that a type (term/word/n-gram) appears in a text. The suffix array data structure makes it convenient to compute the frequency and locations of a substring in a long sequence. This algorithm constructs a suffix array that contains all suffixes, sorted alphabetically. A suffix, also known as a semi-infinite string, is a string that starts at position *i* in the text and continues to the end of the text. Therefore, constructed suffix array shows all possible substrings which are a prefix of suffix. This enables the algorithm to compute the term frequency using overlapping computation. As a result, suffix array can be used to retrieve frequent substring efficiently. The following section depicts using suffix array to compute term frequency and to retrieve frequent substring.

Let input text = "การประกอบการ"

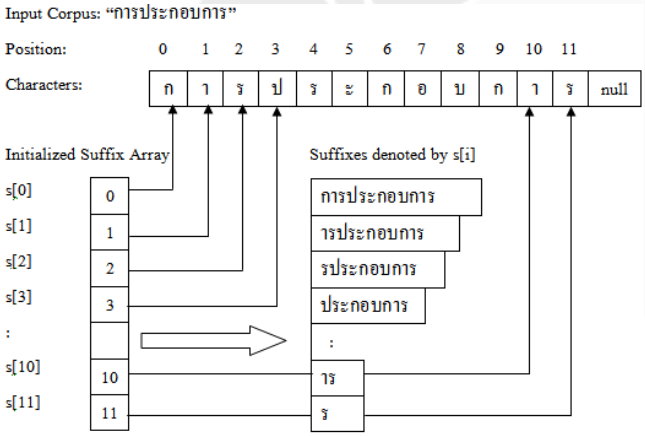


Figure 2. Illustration of a suffix array from input text = "การประกอบการ"

From figure2, the suffixes are enumerated by using suffix array, but elements in the suffix array have not been initialized and sorted. Each element in the suffix array, *s*[*i*], is an integer denoting a suffix or a semi-infinite string, starting at position *i* in the text and extending to the end of the text. The elements in suffix array will then be sorted in alphabetical order for the next process as shown in figure3.

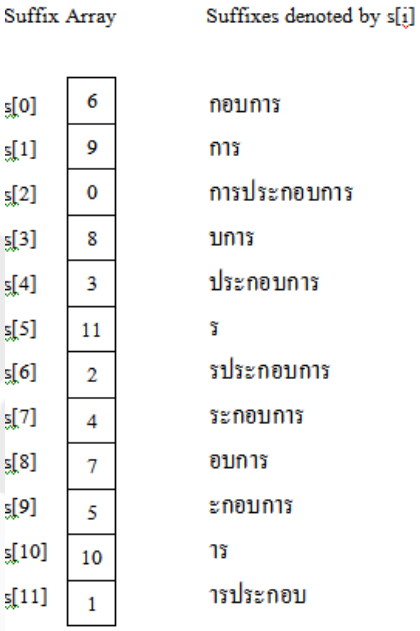


Figure 3. Illustration of a suffix array from figure 2, that has been sorted in alphabetical order.

Additional to using suffix array, algorithms that make use of an auxiliary array for storing LCPs (longest common prefixes) [18] are presented. These algorithms are much faster than the straightforward implementation when the corpus contains long and repeated substrings [19]. This also enables the algorithm to compute the term frequency using overlapping computation as depicted in figure 4 and 5.

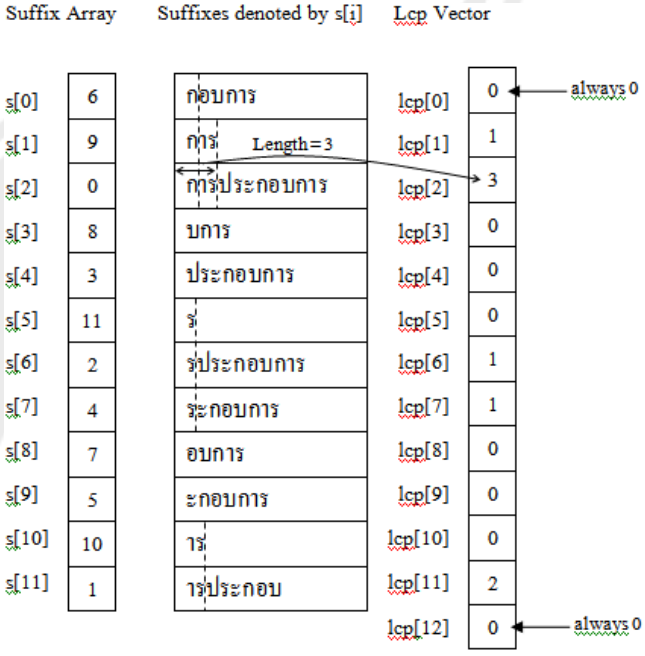


Figure 4. Longest common prefix is vector from suffix array.

LCP-delimited	Terms	LCP-length	Term frequency (TF)
<0,2>	{"ก"}	1	3
<1,2>	{"การ"}	3	2
<5,7>	{"ร"}	1	3
<10.11>	{"าร"}	2	2

Figure 5. All longest common prefix with their length and term frequency

As a result, there are four terms derived from suffix array by using the algorithm to find longest common prefixes.

3.3 Frequent Max substring technique

In this paper, Frequent Max substring technique is proposed as an alternative method to extract Thai keywords [20]. The frequent max substring technique is a substring patterns mining technique used to classify the terms called Frequent Max substring patterns or FM from the non-segmented texts where the word boundary and characteristic are not clearly defined. This technique was first introduced in [6] and has been proposed for indexing un-delimited texts [20] and non-segmented document clustering [21]. The FM refer to all substrings which appear frequently at least the pre-defined frequency and have the maximum length of substring on the given texts, so these terms are likely to be the patterns of interest. In this technique, Frequent Suffix Trie or FST structure is used to enumerate the FM from Thai texts [22]. Frequent Max substring mining technique is based on text mining that describes a process of discovering useful information or knowledge from unstructured texts. We extract the set of FM by using the frequent max substring mining technique. In this technique, the parameter and term frequency are applied to reduce the number of substrings. This method uses two reduction rules: 1) reduction rule using defined frequency to check extracting termination, 2) reduction rule using super-substring definition to reduce the number of substring extracted. In this technique, the algorithm also uses heap data structure to support computation [6]. As a result, this method extracts only the frequent long substrings which contain all frequent substrings from the text.

In order to explain the concept, we will describe the process of extracting frequent max substrings as the keywords by using an example with string S = "การประกอบการ" and the pre-defined frequency = 2

1. We extract *1*-length, *2*-length substrings and so on with their frequencies and positions. Only substrings that occur at least at the pre-defined frequency will be extracted, and these terms are sorted in order of occurring in the texts on the structure.
2. For the second step, we extract the frequent max substring by selecting substrings having no super-substring from the set of the frequent substrings in order to reduce the number of the keywords.

From the process, the FST structure can be shown in figure 6.

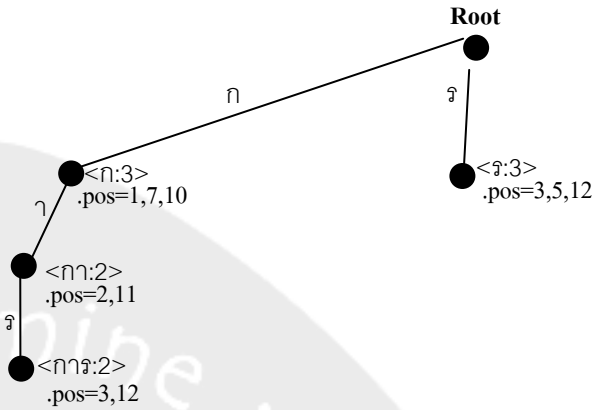


Figure. 6. Shows the FST structure using the efficiency algorithm.

Figure 6 shows the FST structure. The result is FM_{T2} = {<ก : 3>, <ร : 3>, <การ : 2>}

From above three algorithms, we show the keywords of three approaches in Table 2.

Table 2. Shows keywords from three algorithms

The extracted Thai keywords from three algorithms					
Discovering frequent patterns		Suffix array (lcp algorithm)		Frequent max substring	
Terms	TF	Terms	TF	Terms	TF
ก	3	ก	3	ก	3
า	2	การ	2	ร	3
ร	3	ร	3	การ	2
กา	2	าร	2		
าร	2				
การ	2				

From table 2, these three algorithms can be used to retrieve all frequent substrings. However, suffix array techniques provide less number of extracted keywords than Discovering frequent patterns technique, and also Frequent Max substring provides less number of keywords than Suffix array and Discovering frequent patterns techniques. This is because all possible frequent substrings can be

derived from the set of keywords that is extracted by suffix array and Frequent Max substring techniques without information loss.

4. EXPERIMENTS AND DISCUSSION

In this section, we show an experiment to extract keywords using the three algorithms described in section 3: Vilo's algorithm, suffix array and frequent max substring mining techniques as described in the last section. Data used in this experiment is Thai texts found on Thai websites as shown in Table 3.

Table 3. Shows the address of data set

Id	Address of data set
1	http://www.phuketcity.go.th/pkm/index.php?option=com_content&task=view&id=1620&Itemid=97&lang=th_TH
2	http://www.astv-tv.com/news1/viewdata.php?data_id=1001731
3	http://radiothailand.prd.go.th/chonBuri/040newsboard/aspboard_Question.asp?GID=356
4	http://news.mjob.in.th/sport/cat10/news19541/
5	http://news.mumuu.com/sport/page2/
6	http://www.seagames2007.th/th/index.php?option=com_content&task=view&id=1118&Itemid=107
7	http://www.seagames2007.th/th/index.php?option=com_content&task=view&id=1149&Itemid=107
8	http://www.seagames2007.th/th/index.php?option=com_content&task=view&id=1147&Itemid=107
9	http://www.seagames2007.th/th/index.php?option=com_content&task=view&id=1154&Itemid=107
10	http://www.ryt9.com/s/prg/285415/
11	http://www.spokesman.go.th/tape/410721t.txt
12	http://www.freemarketthai.com/tag
13	http://www.matichon.co.th/matichon/view_news.php?newsid=01col01210652&sectionid=0116&day=2009-06-21
14	http://www.naewna.com/news.asp?ID=142987
15	http://203.151.20.17/news_detail.php?newsid=1218170712
16	http://www.welcomethai.com/hotelreservation/news.asp?id=11
17	http://travel.sanook.com/news/news_07858.php
18	http://www.ryt9.com/s/prg/258847/
19	http://news.sanook.com/scoop/scoop_361899.php
20	http://thai.tourismthailand.org/news/release-content-2059.html

In table 4, we show the number of Thai keywords which were extracted from Thai websites by using three algorithms: Vilo's algorithm, Suffix array (lcp algorithm) and frequent max substring mining techniques.

Table 4. Shows number of extracted keywords

The number of extracted keywords using 3 appraches			
	Algorithms		
Text Id	Discovering frequent patterns	Suffix array (lcp algorithm)	Frequent max substring
1	1363	584	156
2	1139	437	145
3	1022	460	204
4	998	345	194
5	3047	1141	312
6	1838	759	329
7	3850	873	247
8	2014	591	360
9	3088	858	209
10	4314	1065	446
11	1997	613	335
12	3400	1104	511
13	3818	918	396
14	4897	1247	525
15	2543	810	475
16	1548	621	409
17	1543	586	398
18	1148	364	256
19	939	382	244
20	5122	1832	775

From table 4, experimental results showed that Frequent max substring technique can extract less number of keywords, but the set of all frequent substrings can still be retrieved from the set of extracted keywords that extracted by using Frequent max substring technique.

5. CONCLUSION

In this paper, an alternative method for extracting Thai keywords is proposed. The proposed approach is based on the analysis of Frequent Max substring that extracts important keywords as long substrings rather than individual words from given texts. As a result, this approach is language-independent. It does not rely on the use of dictionary or language analysis. This technique is referred as Frequent Max substring mining or FM technique that provides less number of Thai keywords that are frequent and highly

distinct from given texts. Experimental studies show that suffix array provides less number of extracted keywords than Vilo's technique. The Frequent Max substring provides less number of keywords than Suffix array and Vilo's techniques. This is because all frequent substrings can still be derived from the set of extracted keywords that are extracted by using Frequent Max substring mining technique without information loss. This shows that Frequent Max substring technique has improved over other algorithms in term of memory storage space in order to support the growth of Thai electronic information.

6. REFERENCES

[1] M. A. Hearst, "Untangling text data mining," Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 3-10, 1999.

[2] A.-H. Tan, "Text Mining: The state of the art and the challenges," Proceedings of the PAKDD Workshop on Knowledge Discovery from Advanced Databases, pp. 65-70, 1999.

[3] V. Sornlertlamvanich, T. Potipiti and T. Charoenporn. Automatic Corpus-Based Thai Word Extraction with the C4.5 Learning Algorithm. In forthcoming Proceedings of COLING 2000.

[4] RattasitSukhahuta and Dan Smith. Information Extraction for Thai Documents. International Journal of Computer Processing of Oriental Languages (IJCPOL), 2001, 14(2):153–172.

[5] ChoochartHaruechaiyasak,PrapassSrichaivattana, SarawootKongyoung and Chaianun Damrongrat. Automatic Thai Keyword Extraction from CategorizedText Corpus, 2008.

[6] T. Chumwatana, , KokWai Wong and Hong Xie, Frequent max substring mining for indexing. International Journal of Computer Science and System Analysis (IJCSSA), India, 2008.

[7] V. Sornlertlamvanich, "Word Segmentation for Thai in Machine Translation System," Machine Translation, National Electronics and Computer Technology Center, Bangkok.

[8] Y. Poovorawan and V. Imarom, "Dictionary-based Thai Syllable Segmentation (in Thai)," 9th Electrical Engineering Conference, 1986.

[9] Y. Thairatananond, "Towords the design of a Thai text syllable analyzer". Master Thesis Asian Institute of Technology.

[10] S. Charnyapornpong , "A Thai Syllable Separation Algorithm. Master Thesis Asian Institute of Technology", 1983.

[11] Theeramunkong, T., Sornlertlamvanich, V., Tanhermhong, T., Chinnan, W., Character-Cluster Based Thai Information Retrieval, Proceedings of the Fifth International Workshop on Information Retrieval with Asian Languages, September 30 - October 20, 2000, Hong Kong, pp.75-80.

[12] ChoochartHaruechaiyasak, SarawootKongyoung and ChaianunDamrongrat, "LearnLexTo: A Machine-Learning Based Word Segmentation for Indexing Thai Texts", In ACM 17th Conference on Information and Knowledge Management, 2008.

[13] C. Kruengkrai and H. Isahara, "A conditional random field framework for thai morphological analysis," Proc. of the Fifth Int. Conf. on Language Resources and Evaluation (LREC-2006), 2006.

[14] H. Arimura, A. Wataki, R. Fujino, and S. Arikawa, "A fast algorithm for discovering optimal string patterns in large text databases," Proceedings of the 9th International Workshop on Algorithmic Learning Theory, pp. 247-261, 1998.

[15] J. Vilo, "Discovering Frequent Patterns from Strings: Department of Computer Science. University of Helsinki, Finland," Technical Report C-1998-9, p. 20, May 1998.

[16] Manber, Udi and Gene Myers. 1990. Suffix arrays: A new method for on-line string searches. In the first Annual ACM-SIAM Symposium on Discrete Algorithms, pages 319–327.

[17] S. J. Puglisi, W. F. Smyth, and A.Turpin : Inverted files versus suffix arrays for locating patters in primary memory. SPIRE 2006, pp 122-133

[18] M. Yamamoto and K. W. Church, "Using Suffix Arrays to Compute Term Frequency and Document Frequency for All Substrings in a Corpus," Computational Linguistics, vol. 27, pp. 1- 30, 2001.

[19] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," In Proceedings of DARPA Speech and Natural Language Workshop, pp. 357-361, 1992

[20] T. Chumwatana, Kok Wai Wong and Hong Xie, "An automatic indexing technique for Thai texts using frequent max substring," *In Eighth International Symposium on Natural Language Processing, 2009 (SNLP '09)*, Bangkok, Thailand, 2009.

[21] T. Chumwatana, Kok Wai Wong and Hong Xie, "Non-segmented Document Clustering Using Self-Organizing Map and Frequent Max Substring Technique" *In16th International Conference on Neural Information Processing (ICONIP 2009)*, Bangkok, Thailand, 2009.

[22] T. Chumwatana, Kok Wai Wong and Hong Xie "THAI TEXT MINING TO SUPPORT WEB SEARCH FOR E-COMMERCE," In *The 7th International Conference on e-Business 2008 (INCEB2008)*, Bangkok, Thailand, 2008.

Using The End-User Computing Satisfaction Instrument to Measure Satisfaction with Web-Based Information Systems

Dedi Rianto Rahadi
Bina Darma University
Jl. A. Yani No 6 Plaju , Palembang, Indonesia
dedi1968@yahoo.com

ABSTRACT

Therefore, it is appropriate to review the measures of user satisfaction with information systems technology, especially in a web-based environment, which accounts for a major component of the end-user computing environment. The objective of this research was to develop and validate an instrument for measuring user satisfaction in a web-based environment.

This research will present significant progress towards keeping the End-User Computing Satisfaction instrument relevant and applicable under the computing environment of Information Era. It will be of significance for both

practical application and theoretical applications. From a practical perspective, this revised End-User Computing Satisfaction (EUCS) instrument can be applied to evaluate end user applications, especially web-based information systems.

Keywords

User satisfaction, end-user computing satisfaction, Web- based systems

1. INTRODUCTION

This paper is to examine critical factors; content, accuracy, format, ease of use, timeliness, satisfaction with system speed and system reliability in End-User Computing Satisfaction (EUCS) that influence most end-users' satisfaction. The research was conducted using a set of questionnaire consist of five factors; content, accuracy, format, ease of use, timeliness, with system speed and system reliability to measure end-users' satisfaction.

End user satisfaction has always been an important component of Information Systems (IS) success. There has been considerable research devoted to establishing a standard user satisfaction instrument since the 1980s (Rita Moore, et al 2007; Pikkarainen et al. 2006; Ali Azadeh et al 2009), when data computing in organizations moved from data processing to end-user computing (EUC) (Doll and Torkzadeh 1988). Doll and Torkzadeh (1988) developed and validated an End-User Computing Satisfaction (EUCS) instrument. It included five components: content, accuracy, format, ease of use, and timeliness. Since the development of the EUCS instrument, there have been significant changes in information technology, especially with the soaring growth of the Internet. For example, widespread use of web technology and rapid increase of Internet-based information systems is evident in the remarkable increase in the number of Internet hosts and web sites. The Internet opened the door to new opportunities for the free flow of information. This information now flows unhindered across local, state, and national boundaries.

Despite the significant changes in the end-user computing environment during the past decade and proliferation of web-based information systems, there has been little research on measurement of user satisfaction with web-based information systems, which is a primary component of end-user computing environment at present. People generally just apply the Doll and Torkzadeh (1988) instrument in their studies to measure the extent of user satisfaction, assuming it is valid and reliable for web-based information systems.

However, there are differences between web-based information systems and traditional corporate information systems. For example, with wide spread use of Internet, access to web-based information systems as well as information has been significantly enhanced. It is much easier to get access to any information that one needs, therefore sufficiency of information provided by information systems may not be an issue for web-based information systems any more. In addition, web-based information systems become more complicated than traditional information systems. More issues other than content, accuracy, format, ease of use and timeliness may be relevant and important in measuring user satisfaction with them. Because of differences between web-based information systems and traditional information systems, it is not appropriate to adopt the EUCS instrument to measure user satisfaction with web -based information systems without examining validity and reliability of the instrument in the specific environment. It is very important to test validity of the instrument in the web-based information system environment. Recently, Pikkarainen *et al*. (2006) have tested the EUCS model in their endeavor to examine user satisfaction with the online banking service. Their results support the validity of three constructs, such as content, ease of use, and accuracy, in the original EUCS model, indicating that the EUCS could be modified and employed to evaluate customer satisfaction with the use of online banking systems.

Consistent with Doll and Torkzadeh's (1988) findings, the authors believe that user involvement should be considered as an independent variable, since end users, like university students, do not have much chance to be involved in the IS development process. However, as the web based information systems has changed its role from a product developer to a service provider, the end user has increasingly interacted with the department as an internal customer. Thus, the authors suggest that the web based information systems service quality should be considered as one of the key attributes of EUCS. Accordingly, the authors propose that a complete measure of

EUCS include items evaluating the following three dimensions: end user satisfaction with computer systems, with information quality provided by the system, and with services provided by the web based information systems.

The purpose of this study is to develop and validate an instrument to measure user satisfaction in the information age. To accomplish this, we first reviewed the literature in the field of user satisfaction measurement. Then we decided to adopt the EUCS instrument by Doll and Torkzadeh (1988) as our starting point. We then checked whether this existing instrument could be used in the new information systems environment. The organization of the paper is as follows. In the next section we provide a review of literature. In the section that follows we present our research methodology. Then we report on the data collection and data analysis, followed by conclusions and discussions of research findings

2. LITERATURE REVIEW

User satisfaction has received considerable attention of researchers since the 1980s as an important surrogate measure of information systems success (Rita Moore, et al 2007; Pikkarainen et al. 2006; Ali Azadeh et al 2009). Several models for measuring user satisfaction were developed, including the user information satisfaction instrument by Ives et al. (1983) and a 12- item EUCS instrument by Doll and Torkzadeh (1988). This instrument included many factors ranging from information quality, systems performance, personal relationship with EDP staff and top management involvement. The End User Information System Satisfaction (EUISS) is probably the most widely used measure of IS success. In this paper, a 12-item instrument that was developed by Doll and Torkzadeh is applied to study EUISS in an Iranian power holding company. This instrument has been vastly accepted in the literature and measures the satisfaction of IS end users in five different dimensions: content, accuracy, format, ease of use and timeliness. However, to date, this instrument has not been tested in an Iranian setting by Ali Azadeh (2009) Limitations of the study involved small sample size (29 valid data) and difficulty of applying the questionnaire. Lei Wang et al. (2007) adopted the instrument by Bailey and Pearson (1983) and examined causal relations of user involvement on system usage and information satisfaction. They concluded that user involvement in the development of information systems enhances both system usage and user‘s satisfaction with the system.

Lei Wang et al. (2007) developed a User Information Satisfaction (UIS) instrument to measure user‘s general satisfaction with the information provided by the data processing group of the organization. Limitations of the study included use of an instrument that was based on the data processing computing environment. The emphasis was on computing tasks that were carried out by the data processing group in an organization. The measuring scale was semantic differential rather than Likert-scale type scaling. Due to the limitations of this study, this instrument is not used as much as the EUCS instrument developed by Doll and Torkzadeh (1988).

Doll and Torkzadeh developed a 12-item EUCS (Figure 1) instrument by contrasting traditional data processing environment and end-user computing environment, which comprised of 5 components: content, accuracy, format, ease of use, and timeliness. Their instrument was regarded as comprehensive, because they reviewed previous work on user satisfaction in their search for a comprehensive list of items. They included measurement of —ease of use,“ which was not included in earlier research. Two global measures of perceived overall satisfaction and success were added to serve as a criterion. The 12 items and the 2 global measures are listed in Appendix 1. The construct was developed with a five point Likert-type scale (1 = almost never; 2 = some of the time; 3 = about half of the time; 4 = most of the time; and 5 = almost always).

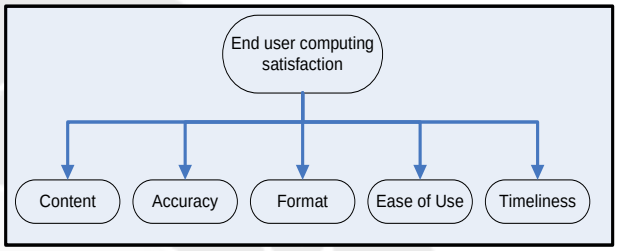


Figure 1. End-User Computing Satisfaction (EUCS) instrument by Doll & Torkzadeh 1988

This research was based on this EUCS instrument by Doll and Torkzadeh because it is a widely used instrument, and has been validated through several confirmatory analyses and construct validity tests. After the exploratory study was completed in 1988, two confirmatory studies with different samples were conducted respectively in 1994 and 1997, which suggested the instrument was valid (Doll et al. 1994; Doll and Xia 1997). A test-retest of reliability of the instrument was conducted in 1991, indicating the instrument was reliable over time (Torkzadeh and Doll 1991). The instrument is widely accepted and adopted in other researches. McHaney and Cronan (1998, 2000) adopted it to examining computer simulation success. McHaney et al. (1999) adopted it in decision support systems research. Roger McHaney et al. (2007) applied it to measure user satisfaction with data warehouse. In the following section we consider the research methodology in detail.

2.1 Research Methodology

Our first step was to examine whether Doll and Torkzadeh instrument (1988) can be used in the new information systems environment, and whether this instrument has to be revised. To do this we decided to follow Doll and Torkzadeh‘s research methodology to conduct the first part of study. We designed questionnaires to survey end users on their satisfaction with web based information systems. In this research, we considered Internet portals to be representative of web-based information systems. Internet portals are widely used among end users of web based information systems. Web portals constitute so me of the most visited sites on the Internet. Digital media measured by the number of unique visitors per day are web portals (www.detik.com). Portals provide a variety of services including web search engines, email, calendar, financial tools, and entertainment. We then followed the steps and measures as Doll and Torkzadeh (1988) did to analyze the data to examine the validity and reliability of the instrument. In the sections that follow we provide information on the pilot study, data collection and analysis.

2.2 Pilot Study

Our pilot study was conducted with a sample of 250 students in univesity. The construct validity was examined with correlations between total score and each item scores. To avoid spurious part-whole correlation, the total score was corrected by subtracting the item score before examining the correlation. Criterion-related validity was also examined using two-item global criterion including "Is the system successful ?" " and "Are you satisfied with the system ?" In order to analyze the pilot study data, we assessed the item-total correlation and criterion-related correlation.

Table 1. Descristipve statistics of research sample

Gender	**Male**		**Female**			**Total**
	175 (70%)		75(30%)			250
Work Experience	**< 1 year**	**1–3 years**	**4-6 years**	**7-10 years**	**>10 years**	
	135 (54%)	85 (34%)	18 (7,2%)	12 (4,8%)	-	250
Location (multiplechoice)	**Home**		**Work**		**School**	
	250 (40%)		138 (22%)		231 (37%)	619

3. SAMPLE

The sample for this study included students at a large mid-Bina Darma university at Palembang, Indonesia. Although the sample for this study was collected at a higher education institution, it well represents the end user population, because some students were full-time students while others were part time students from a variety of industries and management levels. Table 1 provides some descriptive statistics.

3.1 Data Collection

According to Doll and Torkzadeh (1988) there are five components of user satisfaction with information systems: content, accuracy, format, ease of use, and timeliness. This research conducted a survey of 250 end users about their satisfaction/ dissatisfaction with Internet portals. A list of questionnaire questions is provid ed in Appendix 1 .

Data were collected in classes at a large Bina Darma university in Indonesia. Students were told that participation in this study was voluntary and anonymous. No personal identity information was collected during the survey. Hard copy questionnaires were distributed in class. It took 10 minutes to complete the survey.

3.2 Data Analysis

In this study, we followed the methodology used by Doll and Torkzadeh (1988) to analyze the data. We analyzed the construct validity, examined criterion-related validity, and reliability. We evaluated the construct validity and the constructs in the EUCS instrument.

3.3 Factor Analysis

In conducting the factor analysis, we exp ected the factors (questions in our study) to load on constructs originally identified by the earlier study. The Principal Components Analysis (PCA) was used as the extraction technique and varimax was used as a method of rotation. Table 2 is the factor matrix of the 12-item instrument. We took the threshold value of 0.7 for factor loading criterion.

Table. 2 Rotated Factor Matric of the 12 -Item Instrument

Item/ Question Code	Content	Accuracy	Format	Ease of Use	Timeliness
C1	0.849				
C2	0.824				
C3	0.755				
C4	0.701				
A1		0.811			
A2		0.776			
F1			0.836		
F2			0.675*		
E1				0.867	
E2				0.897	
T1					0.806
T2					0.706

* The loading of question F2 is 0.675, which is close to 0.7, therefore, we decided to keep it in the instrument.

As we can see from the factor matrix, the primary loadings for the five factors are well above 0.7 while the factor loading for the question F2 is very close to 0.7 (for question F2 see Appendix 1). Therefore we keep all the factors as they are in the instrument.

Next we conducted item-total correlation as well as criterion related correlation. Following Doll and Torkzadeh's procedure, we examined the correlation of score of each item with the total score of all questions. To avoid the spurious part-whole correlation, we subtracted each item score from the total score before conducting the correlation, therefore we conducted correlation of each item with the total of rest 11 items. Table 3 lists the result of the correlation assessment. According to Doll and Torkzadeh, there is no accepted standard of cut off threshold, therefore we took the same cutoff value of 0.5 as they did in their study.

Table 3. Item –Total Correlation

Factor	Correlation Coefficient	Alpha
C1	0.638	<.0001
C2	0.679	<.0001
C3	0.605	<.0001
C4	0.223	<.0001
A1	0.637	<.0001
A2	0.670	<.0001
F1	0.600	<.0001
F2	0.650	<.0001
E1	0.585	<.0001
E2	0.580	<.0001

T1	0.553	<.0001
T2	0.481	<.0001

As we find out from the correlation coefficient, all questions coefficient is above the threshold of 0.5, except for the question C4 which is well below the threshold.

In conducting the criterion-related validity analysis, we examined the correlation of each item with the score of two global satisfaction criteria G1 and G2 in Appendix 1, and questions 13 and 14 in Appendix 2. As Doll and Torkzadeh, we assumed that the two global measures of end-user satisfaction to be valid. Table 4 is the result of item-criterion correlation. The cut off threshold is 0.4 as Doll and Torkzadeh did in their research.

Table 4. Item Criterion Correltion

Item	Correlation Coefficient	Alpha
C1	0.431	<.0001
C2	0.506	<.0001
C3	0.436	<.0001
C4	0.139	0.0113
A1	0.519	<.0001
A2	0.585	<.0001
F1	0.498	<.0001
F2	0.463	<.0001
E1	0.533	<.0001
E2	0.536	<.0001
T1	0.482	<.0001
T2	0.523	<.0001

As in the item-total correlation, all factors have correlation coefficients of greater than 0.4 threshold value except for the question C4 has correlation coefficient of 0.139, well below the threshold. Therefore all questions other than question C4 — "Does the system provide sufficient information?" are valid. And we also observed that components of satisfaction as identified by Doll and Torkzadeh are still relevant for users of web-based information systems. Due to the results of data analysis, we dropped the question C4, — "Does the system provide sufficient information?"

4. DISCUSSION

This research will present significant progress towards keeping the End-User Computing Satisfaction instrument relevant and applicable under the computing environment of Information Era. It will be of significance for both practical application and theoretical applications. From a practical perspective, this revised End-User Computing Satisfaction (EUCS) instrument can be applied to evaluate end user applications, especially web-based information systems.

End user satisfaction has always been an important component of Information Systems (IS) success. This is also true for online applications, including online shopping systems where in addition to being a customer, the shoppers play the role of end users. Shoppers may not come back to or make a purchase on a website if they have an unsatisfactory experience. In this research, we focus on this aspect of online shopping by examining shoppers' experiences as end users by Liu (2007).

This study provides several implications for researchers. 1), should attempt to identify additional components of satisfaction that are specific to a web-based environment. Some components that could be relevant are privacy and security. Although we consider our sample to be appropriatecontinuing research efforts should be made to improve the combined measure of this study. 2), there is still a need for a more reliable measure of IS service quality with merged two sets of items, pertaining to Doll and Tokzadeh's (1988) EUCS and Kettinger and Lee's (1997) IS-adapted SERVQUAL repectively, into one. Kettinger and Lee (1997) find an alternative approach to alleviating those problems associated with SERVQUAL. They have undertaken further refinement of SERVQUAL to improve its practical value to IS managers and have suggested a 13-item IS-Adapted SERVQUAL. This IS adapted SERVQUAL includes four of the five original SERVQUAL dimensions: reliability, responsiveness, assurance, and empathy.

This approach assumed that the two instruments were well-developed and could precisely measure the factors that they were supposed to measure. However, as mentioned in the literature review section, there has been a controversy over whether the factor structure of the IS-Adapted SERVQUAL is valid. Moreover, the results of this study suggest that some items included in the aforementioned two original measures are not significant in the combined measure. Therefore, the authors strongly encourage research efforts to incorporate more items from the original SERVQUAL instrument into the new measure of this study. In addition, the authors encourage future research to utilize at least three items that assess the following two dimensions, such as "format" and "accuracy", which were represented by only two items in Doll and Tokzadeh's (1988) EUCS measure.

In addition to overall user satisfaction assessment with information systems, it can also used to compare end-user satisfaction with different components of end-user computing task. Last but not least, it can be used to compare different information systems that perform the same functions. From a theoretical perspective, by dividing the End-User Computing Satisfaction construct into separate components, this research provides a valuable theoretical framework to enable more precise research with the components.

This research might begin with a focus on expectations. Little is known about the expectations of IS service customers and users. A better understanding of expectations would help increase our understanding of the service quality construct. Research should focus on different types of study populations such as internal versus

external customers, purchasers versus end users, and differences based on varying levels of experience with the IS function. Such research would constitute an important step in the development of an improved measure of IS

5. CONCLUSIONS

With ever-changing technology and significantly different end-user computing environment, it is necessary to develop and validate an instrument to measure user satisfaction with information systems in the information age. In this study, we have taken the first step towards fulfilling this objective. Our starting point was the end-user computing satisfaction instrument developed by Doll and Torkzadeh (1988). We retested this instrument to measure satisfaction in a web-based

environment. We found that with minor revisions the new instrument provides a valid measure of user satisfaction. Every study has its limitations, and this one is no exception. Our study focused on a specific kind of web-based information system, Internet portals. The second limitation arises from the components of satisfaction. We did not identify and test for any additional components of user satisfaction. It is possible, even likely, that there are some other components of satisfaction such privacy that are unique to web-based systems that have not been considered in this study. Some researchers may consider the use of students to be a limitation. But, we believe that students are a representative part of the general population of Internet users. There are a number of avenues of future research. As we mentioned in the limitations, we measured satisfaction using established measures or components. The five components we used were content, accuracy, format, ease of use, and timeliness. By doing so, future research may develop a measure with better validity and reliability. Additional research could investigate the components of satisfaction in web-based systems in a work-related environment. While we conducted our study using Internet portals, future research can extend our study to examine satisfaction with other web-based systems. In sum, the authors suggest the IS research community to keep moving toward generating a reliable and valid instrument for measuring EUCS that can fit the constantly changing environment of IS. Such an updated instrument will provide both academicians and practitioners with better insight into the nature and determinants of EUCS. We have successfully revised and tested an instrument to measure user satisfaction with web-based systems. We believe that this is a step in the right direction, and we recommend additional research.

6. REFERENCES

[1] Ali Azadeh , Mohamad Sadegh Sangari , Mohsen Jafari Songhori, An empirical study of the end-user satisfaction with information systems using the Doll and Torkzadeh instrument, International Journal of Business Information Systems, Volume 4, Number 3 / 2009, Pages: 324 – 339

[2] Doll, William J. , Deng, Xiaodong , Raghunathan, T. S. , Torkzadeh, Gholamreza and Xia, Weidong, The Meaning and Measurement of User Satisfaction: A Multigroup Invariance Analysis of the End-User Computing Satisfaction Instrument, Journal of Management Information Systems , Vol. 21 No. 1, Summer 2004 pp. 227 - 262

[3] Doll, W. J. and Torkzadeh, G. .The Measurement of End-User Computing Satisfaction,. MIS Quarterly (12:2), June 1988, pp. 259-274.

[4] Doll, W. J., Xia, W. and Torkzadeh, G. .A Confirmatory Factor Analysis of the End-User Computing Satisfaction Instrument,. MIS Quarterly, December 1994, pp. 453-461.

[5] Ives, B., Olson, M. H. and Baroudi, J. J. . The Measurement of User Information Satisfaction,. Communications of the ACM (26:10), October 1983, pp. 785-793.

[6] Pikkarainen, Tero Pikkarainen, Heikki Karjaluoto, Seppo Pahnila, The measurement of end-user computing satisfaction of online banking services: empirical evidence from Finland, International Journal of Bank Marketing, 2006 vol 24 isuue 3 page 158-172

[7] Kettinger, W., Lee, C. , "Pragmatic perspectives on the measurement of information systems service quality", *MIS Quarterly*, 1997, Vol. 21 No.2, pp.223-40.

[8] Lei Wang; Youmin Xi; Huang, W.W., A Validation of End-User Computing Satisfaction Instrument in Group Decision Support Systems, Wireless Communications, Networking and Mobile Computing, 2007. WiCom 2007. International Conference on, Volume , Issue , 21-25 Sept. 2007 Page(s):6031 – 6034

[9] Liu, Chung-Tzer; Guo, Yi Maggie Validating the End-User Computing Satisfaction Instrument for Online Shopping Systems, 2007, Journal of Organizational and End User Computing, Vol. 20, Issue, pages 74-96

[10] McHaney, R. and Cronan, T.P. .Computer Simulation Success: On the Use of the End-User Computing Satisfaction Instrument: A Comment,. Decision Sciences (29:2), Spring 1998, pp. 525-536.

[11] McHaney, R. and Cronan, T.P. .Toward an empirical understanding of computer simulation implementation success,. Information and Management (37), 2000, pp. 135-151.

[12] McHaney, R. Hightower, R. and White D. .EUCS test-retest reliability in representational model decision support systems., Information and Management (36), 1999, pp. 109-119.

[13] Rita Moore, Mary Jo Jackson, Ronald B. Wilkes, End-user computing strategy: an examination of its impact on end-user satisfaction, Academy of Strategic Management Journal, Annual, 2007

[14] Roger McHaney [1] Timothy Paul Cronan, Computer Simulation Success: On the Use of the End-User Computing Satisfaction Instrument: A Comment, Decision Sciences, Published Online: 7 Jun 2007, Volume 29 Issue 2, Pages 525-535

[15] Torkzadeh, G. and Doll, W. .Test-Retest Reliability of the End-User Computing Satisfaction Instrument., Decision Sciences (22:1), winter 1991, pp. 26-37.

Batik Image Classification Using Log-Gabor and Generalized Hough Transform Features

Laksmita Rahadianti
Faculty of Computer Science
University of Indonesia
Depok INDONESIA
lara50@ui.ac.id

Hadaiq R. Sanabila
Faculty of Computer Science
University of Indonesia
Depok INDONESIA
hadaiq.rolis@ui.ac.id

Ruli Manurung
Faculty of Computer Science
University of Indonesia
Depok INDONESIA
maruli@cs.ui.ac.id

Aniati Murni
Faculty of Computer Science
University of Indonesia
Depok INDONESIA
aniati@cs.ui.ac.id

ABSTRACT
Batik is a decorative cloth used widely in the Indonesian culture. Indonesian batik also is very much diverse and consists of many different patterns, colors, and textures. In the course of further research on Indonesian culture heritage, we are trying to develop an automatic batik classification system using the characteristics and features contained in each batik cloth. Using a self-proposed taxonomy of batik patterns, we conduct experiments to differentiate various batik patterns using two well-known image processing features, namely the log-Gabor feature vector and the Generalized Hough Transform. Building on previous work, we attempt to account for variations in scale and orientation. Our experiments are able to show good accuracy, particularly using the Generalized Hough Transform. For the log-Gabor filter, further experimentation is required to determine the optimal configuration of scale and orientation parameters in constructing the filter bank.

Keywords
Batik, image classification, Generalized Hough Transform, Log Gabor Filter

1. INTRODUCTION
Decorative cloths are present all over the world in many communities. These cloths have various functions in each community and area, but every cloth is unique, differing in color, pattern, and texture. Batik is a decorative cloth used widely in Indonesian culture. Indonesian batik is also very much diverse and consists of many different patterns, colors, and textures [1].

In the course of further research on Indonesian culture heritage, it has come to attention the viability to automate the batik differentiation process using the characteristics contained in each batik cloth. Each batik cloth has specific characteristics in pattern, texture, and color. These characteristics should be able to be used as a factor in distinguishing these batik cloths from one another. This automated process is done using digital batik images and computing the images mathematically. The development of this system has been initiated by previous work conducted at the Faculty of Computer Science, University of Indonesia [2].

In an attempt to establish this automated batik recognition system, a number of follow-up studies have been conducted in order to determine the ability of the batik image characteristics contained in digital images to represent the batik cloth. The characteristics of each batik image are represented using various computations, such as color histograms to collect information on color, or the log-Gabor filter and Hough transform to detect patterns and textures [3, 4].

The ability to differentiate batik cloths is still limited to those with sufficient knowledge of batik history and predefined batik patterns. There are many references regarding the different patterns and textures that build the Indonesian batik pattern library with different approaches to the structure of the pattern taxonomy. Batik scholars and experts define the batik pattern taxonomy according to their respective opinion [1,5].

In this paper we propose our own version of batik pattern taxonomy that is suitable for the computation and processing of digital images of batik cloths (Section 2). We then conduct experiments to recognize batik patterns according to the taxonomy. We then present the results of batik pattern classification using log-Gabor (Section 3) and Hough Transform (Section 4) features as the distinguisher.

2. PROPOSED BATIK TAXONOMY
As mentioned in Section 1, the organization of batik patterns is until this day different according to each expert. Previously known batik pattern definitions were separated according to batik origin and color [5], or simply divided directly into known batik patterns [6]. For our automatic batik recognizer, we use a self-proposed taxonomy, which was initially built from a database composed of templates of well-known batik patterns that were already used in the previous research conducted [2]. This database was populated by a number of batik patterns identified according to a former batik pattern taxonomy [6]. In the course of adding more images and patterns to the batik database, data collection was conducted; mainly from the same source, but furthermore from extensive data exploring also.

This process left us with a limited number of distinguishable batik patterns projected to be suitable for further implementation of the automated batik recognizer project. This taxonomy is hoped to be the simplest and most suitable organization of patterns to be recognized using the mathematical computation and feature extraction of digital batik images. Nevertheless is should be noted

that this taxonomy is still incomplete and has plenty of room for more exploration.

Batik images can be labeled with characteristics it contains. Each batik cloth can contain more than one primitive batik pattern, each of which we aim to be able to recognize using the automated batik recognizer we intend to build. Aside from that, each batik cloth contains color information. We define Indonesian batiks to be divided into two classes of color, namely Sogan and non-Sogan. Sogan batiks are batik cloths that are coloured broken white, black, brown, dark red, indigo and crème [5]. Non-sogan batiks are batik cloths with any other color aside from those stated before.

We define this taxonomy dividing the possible batik patterns based on three main attributes: the area of origin, the color family, and the batik motif/pattern. The possible values can be seen in Table 1.

Table 1. Proposed Batik Taxonomy.

Color Family	Area of Origin	Batik Motif/Pattern
Sogan	Yogya/Solo	Cakar
		Truntum
		Ceplok
		Kawung
		Grompol
		Parang
		Nitik
	Solo	Tirtotejo
non-Sogan	Cirebon	Kumpeni
		Megamendung

3. CLASSIFICATION USING LOG-GABOR FILTER FEATURES

The visual content of a two-dimensional image is, at its lowest level, represented as the colour and intensity of each pixel. However, to be able to classify such images based on visual content similarity, these pixel values must first be transformed into a feature space which forms a more appropriate higher-level representation, upon which we can compute similarity measures that agree with human perception.

One very popular approach is to express the image by transforming it from the spatial into the frequency domain, e.g. using the Fourier transform. This approach is based on the principle of expressing periodic functions using the sum of sine and cosine functions, where the periodic function is represented as an index of the periods and amplitudes of the sine and cosine functions used. Although a periodic function is linear and one-dimensional, the transformation can be applied to images as well. However, this results in an issue of periodicity. In order to localize the information into smaller pieces building up to represent the

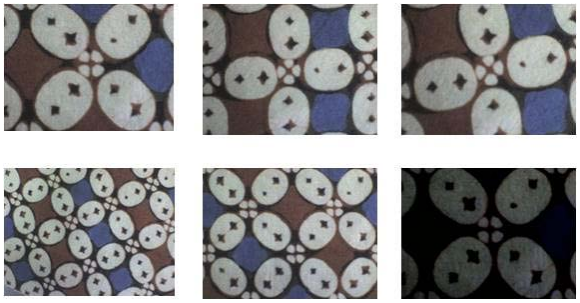


Figure 1. Variations of Batik Pattern Images.

frequency information, we use wavelets, i.e. small portions of a wave used initially summing up to the function intended. Each wavelet used can be dilated and shifted to meet the specific image. The final result would be the values of dilation and shifts on each wavelet involved.

3.1 Log-Gabor filters

In recognizing batik patterns from digital images the main problem is in the variation of scale and orientation of certain patterns. Other than that there may also be the issue of brightness and lighting. The various possibilities can be seen in Figure 1. These problems can be overcome with Gabor filters, which are invariant to scale as well as orientation. Gabor filters have been widely used for texture segmentation, classification, face recognition, and content-based image retrieval systems, such as in [9].

Gabor filters are based on Gabor functions. A two-dimensional Gabor function $g(x,y)$ can be written as [11]:

$$g(x,y) = \left(\frac{1}{2\pi\sigma_x\sigma_y}\right)\exp\left[-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}\right)+2\pi jWx\right]$$

Let $g(x,y)$ be a mother wavelet, a class of self-similar functions or children wavelets, referred to as Gabor wavelets, can be obtained by dilations and rotations of $g(x,y)$ through the generating function:

$$g_{mn}(x,y) = a^{-m}g(x',y'),\ m,n \text{ are integers}$$
$$x' = a^{-m}(x\cos\theta + y\sin\theta)$$
$$y' = a^{-m}(-x\sin\theta + y\cos\theta)$$

where $a > 1$, $\theta = n\pi/K$, n = 1,2, … S, S is the total number of scales, and K is the total number of orientations.

The Gabor wavelet transform for a given image $I(x,y)$ is thus defined as

$$W_{mn}(x,y) = \int (x_1,y_1)g_{mn}*(x-x_1,y-y_1)dx_1dy_1$$

where * indicates the complex conjugate. The mean μ_{mn} and the standard deviation σ_{mn} of the magnitude of the transform coefficients are:

$$\mu_{mn} = \iint |W_{mn}(xy)| dxdy$$

$$\sigma_{mn} = \sqrt{\iint \left(|W_{mn}(xy)| - \mu_{mn}\right)^2 dxdy}\,.$$

A feature vector is constructed using μ_{mn} and σ_{mn} as feature components.

A variant of Gabor filter, namely log-Gabor filter, was proposed by Field [8], who suggests that log-Gabor filters, which have extended tail at the high frequency end, should be able to encode natural images better than Gabor filters. Gabor filters over-represent the low frequency and under-represent the high frequency in any encoding.

A frequency response of a log-Gabor filter is defined as

$$G(f) = \exp\left(-\left[\log(f / f_o)\right]^2 / 2\left[\log(\sigma / f_o)\right]^2\right)$$

where f_o is the centre frequency of the filter and σ is a constant [8]. A log response of a log-Gabor filter is symmetric on a log axis, which is the standard method for representing the spatial-frequency responses. Conversely, a Gabor filter fails to capture symmetry on log axis. Therefore, Field suggests that log-Gabor filters may provide a better description than Gabor filters.

3.2 Experimental setup

Before performing the log-Gabor filter on the images, they are first converted into binary images, to mitigate the factor of different lighting conditions.

The log-Gabor feature vector is constructed by convolving the image with a filter bank consisting of Gabor filters at a certain number of angles and scales. This is intended to cover the various scales and orientations possible to occur in a certain image. The optimal number of scales and orientations needed to process batik patterns is to be explored. In this preliminary experiment, we try to find a suitable combination of these values.

In our previous attempts at batik pattern classification [2], we treat each image in our batik database as an individual image, and we retrieve the image with the most similar feature vector. However, in our current experiment we define the classification task to be the correct identification of the batik pattern, which is actually represented in the database as a set of various images. Thus, to determine the batik pattern of a query image, we first compute the distance of its log-Gabor feature vector to each feature vector in the database, and then compute its average distance to each set of exemplar images belonging to each pattern. In other words, the matching is not done simply to an individual image, but to the entire set of images from a certain pattern.

In our experimental setup we intend to differentiate between three patterns, namely the *Megamendung* pattern from Cirebon, the *Kawung* pattern from Yogyakarta, and the *Udan Liris* pattern from Yogyakarta (see Figure 2). These patterns were randomly chosen from the patterns in our current taxonomy. Each pattern is represented by the various log-Gabor feature vectors of each exemplar image.

Figure 2. Examples of images used.

For testing, we prepare five different images from each pattern cluster as query images. To compare the patterns, the log-Gabor features of a query image is compared to all images of a pattern set, and the value is averaged between the set. This isdone for each image pattern set. The classifier assigns the pattern based on the pattern set with the closest average distance. The final result was calculated using accuracy of classification of the queries for each pattern.

In our experiment, we vary the scale and orientation parameters of the Gabor filter bank used for convolution. The proposed numbers are 4 for scale and 6 for orientation [2], but we also explore the possibilities of using higher numbers, namely 4, 50, and 100 for scale and 6, 18, and 36 for orientation. The results are shown in Table 2.

3.3 Analysis of results

The results show that the highest possible accuracy achieved for pattern 1 (*Megamendung*) is 60%. This is reached with 18 orientation values and 50 or 100 scales. As for pattern 2 (*Kawung*), 100% accuracy is achieved using 36 values of orientation and 50 or 100 scales. An accuracy of 100% is also achieved with pattern 3 (*Udan Liris*), but at a scale of 4 and orientation of 18 or 36. Analysing these results, there does not seem to be a clear trend that can be used to predict the optimal number of scales and orientation for recognizing batik patterns. Indeed, designing Gabor filter banks for a specific task is a difficult process, and one that warrants further exploration. Another observation is that the highest accuracy achieved with pattern 1 (Megamendung) is not as high as with the other patterns. Looking at Figure 1, we can see that the exemplar images are of different cloths that belong to the same set, whereas the exemplar images for patterns 2 and 3 are actually scale and rotation variations of the same cloth.

Table 2. Classification results using log-Gabor Features.

ACCURACY			Orientation		
			6	18	36
Scale	4	Pattern 1	17%	0%	0%
		Pattern 2	80%	80%	20%
		Pattern 3	0%	100%	100%
		Average	31%	60%	40%
	50	Pattern 1	0%	60%	0%
		Pattern 2	20%	0%	100%
		Pattern 3	80%	0%	0%
		Average	33%	20%	33%
	100	Pattern 1	0%	60%	0%
		Pattern 2	20%	0%	100%
		Pattern 3	80%	0%	0%
		Average	33%	20%	33%

4. CLASSIFICATION USING HOUGH TRANSFORM FEATURES

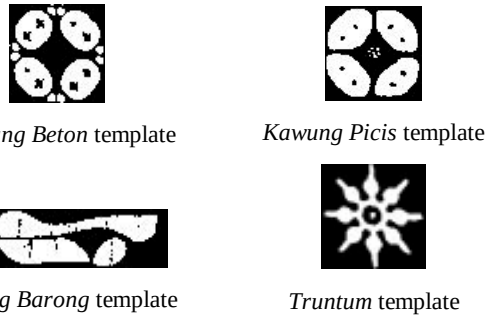
Template matching is a well known method for recognizing a pattern contained within an image. This is a process to match a template to an image, where the template is an image or a sub-image that contains the pattern we are trying to find. Although this method is quite simple and easy to implement, it is difficult to obtain a satisfactory result. This is mainly caused by variations of scale and rotation of the template, as it is stored as a discrete set of points.

4.1 Generalized Hough Transform

The Hough Transform is a feature extraction technique widely used in image processing and analysis. It is a robust template matching method that exhibits scale and rotation invariance. The classical Hough transform uses parametric equations of shapes to effect the transformation from image space into parameter space. However, in the case of arbitrary shapes such as batik motifs, we do not have a simple analytic formula to describe its boundary. Therefore the Generalized Hough Transform can be used to generalize the boundary of template shape in the images.

It uses a voting procedure in order to find the reference points which indicate the template image presence in the image to be recognized. The voting procedure is presented in parameter space and the calculation procedure is recorded in an accumulator array.

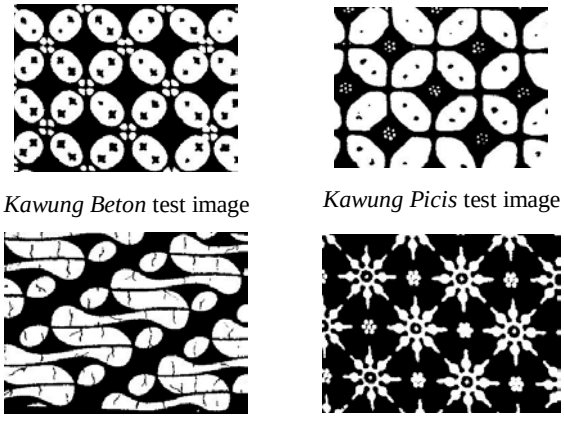
The Generalized Hough Transform uses the boundary curve (edge point) of images to be recognized and template images as the identification parameter (see Figures 3 and 4). The information between the edge point and reference point are stored in a lookup table called the R-Table. The R-Table is constructed during the training phase and uses the gradient of edge pixel as additional information.



Kawung Beton template *Kawung Picis* template

Parang Barong template *Truntum* template

Figure 3. Examples of template images.



Kawung Beton test image *Kawung Picis* test image

Parang Barong test image *Truntum* test image

Figure 4. Examples of batik fabric test images.

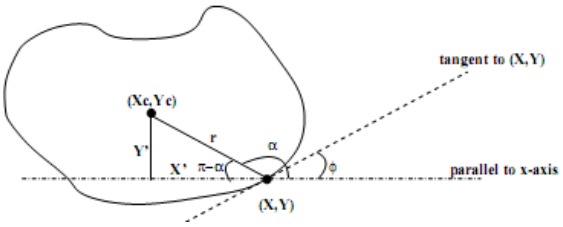


Figure 5. The illustration of *r* and *α* calculation.

To analyze the boundary curve, Generalized Hough Transform defines the following parameters for arbitrary shape: $a = \{y, s, \theta\}$, where $y=(x_r, y_r)$ is a reference coordinate for the origin shape, $s=(s_x, s_y)$ is a scale factor, and θ is its orientation. The reference origin location, y, is described in terms of table possible edge pixel orientations. The calculation of additional parameters s and θ is attained by straight transformation of the table.

For example, given the arbitrary shape in Figure 5, the reference point is at coordinate (x_c,y_c), and α describes the direction between the edge point and the reference point, and r is the distance between the edge point and the reference point.

The R-Table has a main role in the Generalized Hough transform as the modification of the equation of curve shape in an image. The construction of the R-Table is conducted by examining the edge

points and reference points. More specifically, the R-table construction is accomplished as follows:

1. Define the reference point (x_c,y_c) somewhere inside the shape.
2. For each edge point (x,y) in the edge of the shape, find the two parameters:

$$r = \sqrt{(x - x_c)^2 + (y - y_c)^2}$$

$$\alpha = \tan^{-1}\left(\frac{y - y_c}{x - x_c}\right)$$

   and the gradient direction.
3. Store the pair (r,α) of the reference point (x_c,y_c) as a function of gradient (i.e. build the R-Table)

4.2 Experimental setup

In an initial experiment, we employed the Generalized Hough Transform to detect the presence of batik motifs without variation of scale and rotation [4]. The results showed that this method is quite promising to detect such arbitrary shapes. In this experiment, we test the performance of the Generalized Hough Transform to detect batik motifs in various scale and orientation configurations, and attempt to classify them based on motif. We examine the presence of four batik motif templates in their corresponding test images, i.e. *Kawung Beton*, *Kawung Picis*, *Parang Barong*, and *Truntum*, which can be seen in Figures 3 and 4. For each test image, we apply the Generalized Hough Transform using each of the four templates, and detect the local maxima arising in the final accumulator array. The test image is classified according to the template that yields the highest value.

Furthermore, for each test image, we create variations of two scales and four orientations, e.g. 45^0, 135^0, 225^0, 315^0 variance. Examples for the *Kawung Beton* motif can be seen in Figure 6.

4.3 Analysis of results

The results are shown in Table 3. For each batik motif, we compute the accuracy of the system in terms of the percentage of times the motif of an image is correctly determined, averaging over all variations of scale and orientation. Based on the results, we can observe that the Generalized Hough Transform was successful in tracing the location and presence of the batik motif templates in the test images. This resulted in accurate classification every time, and proves the robustness of the Generalized Hough Transform method for dealing with scale and rotation variations in batik motif recognition.

Table 3. Recognition rate of all batik test images, averaged over all scale and orientation variations.

Batik motif	Accuracy
Kawung Beton	100%
Kawung Picis	100%
Parang Barong	100%
Truntum	100%

5. CONCLUSION

The results in Section 3.3 show that the log-Gabor filters, whilst demonstrating the ability to successfully classify batik patterns under certain configurations, still warrant further exploration to discover the optimal factors of scale and orientation in constructing the filter bank. Moreover, when a batik pattern is represented in the database by a set of diverse images, the ability to correctly classify a query image may be compromised. As we can assume, the log-Gabor computations applied to batik images must be altered to be able to process the batik images correctly. This is a subject of our further experimentation.

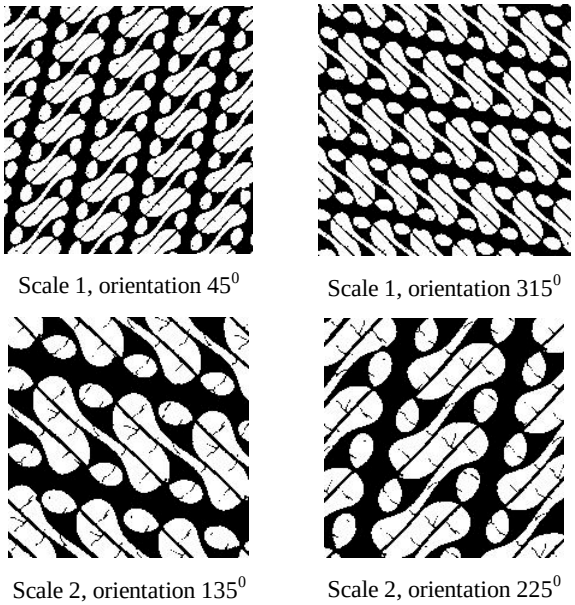


Figure 6. Example test images in various scale and

The results of the generalized Hough transform experiment, on the other hand, results in an accuracy of 100% in an attempt to prove the validity of the taxonomy proposed. However, the weakness of this method is its very expensive time and space complexity. Therefore, further effort must be made to explore and modify the Generalized Hough Transform in order to reduce the resource requirements. One such effort is exploring Multiresolution Hough Transform [10]. Another alternative is to explore the possible combination of the two features presented inthis paper, the log-Gabor filter features and the Generalized Hough Transform features. Our main goal is to obtain a robust and cost-efficient method to detect the batik motifs in batik fabric images.

As a continuity of this research, the long term purpose of this whole study is to establish a reliable automated batik pattern recognizer that can be used to distinguish batik patterns from one another. A prototype system has previously been developed, resulting in a mobile application using web services which enacts a content-based image retrieval system that can retrieve batik images according to a query image [2].

Using further exploration and experimenting on many other features a batik image can contain, we intend to refine the design of the said system according to the existing batik patterns defined in our taxonomy. The new batik pattern recognizer will be designed to be able to detect the patterns contained in each batik image. If a batik image contains more than one known pattern, using the established methods, the batik pattern recognizer will be able to

distinguish each of those batik patterns. This system will then be able to retrieve relevant information regarding the said image.

6. REFERENCES

[1] van Roojen, P. 2001. Batik Design. Amsterdam: The Pepin Press.

[2] Margaretha, E., Azurat, A., Manurung, R., and Murni, A. 2009. Content-based information retrieval system for batik application. Technical Report. University of Indonesia.

[3] Rahadianti, L., Manurung, R., and Murni, A. 2009. Clustering batik images based on log-gabor and colour histogram features. In Proceedings of the International Conference on Advanced Computer Science and Information Systems, Depok, Dec. 2009, pp.85-90, ISSN 2086-1796.

[4] Sanabila, H.R. and Manurung, R. 2009. Recognition of Batik Motifs using Generalized Hough Transform. In Proceedings of the International Conference on Advanced Computer Science and Information Systems, Depok, Dec. 2009, pp.79-84, ISSN 2086-1796.

[5] Djoemena, N. S. 1986. Ungkapan Sehelai Batik, Its Mystery and Its Meaning. Jakarta: Penerbit Djambatan.

[6] Harimawan, R. K. 2005. Mengenal Motif dan Seni Batik Tradisional. Yogyakarta: Tjokrosuharto Arts and Crafts.

[7] McAndrew, A. 2004. Introduction to Digital Image Processing with Matlab. Melbourne: Thompson Couse Technology.

[8] Field, D.J. 1987. Relations Between the Statistics of Natural Images and the Response Properties of Cortical Cells, Journal of The Optical Society of America A, Vol 4, No. 12, December 1987. pp 2379-2394

[9] Andrysiak, T. and Choras, M. 2005. Image Retrieval Based On Hierarchical Gabor Filters, Int. J. Appl. Math. Comput. Sci., vol. 15, no. 4, pp. 471–480, 2005.

[10] Atiquzzaman, M. 1992. Multiresolution Hough Transform - An Efficient Method of Detecting Patterns in Images. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.14, Issue 11 (Nov.1992) , 1090 – 1095.

[11] Manjunath, B.S. and Ma, W.Y. 1996. Texture Features for Browsing and Retrieval of Image Data, IEEE Trans. on Pattern Analysis And Machine Intelligence, vol. 18, no. 8, pp. 837-842, Aug. 1996

Burrows Wheeler Compression Algorithm (BWCA) in Lossless Image Compression

Elfitrin Syahrul
Université de Bourgogne
University of Gunadarma
elfitrin@staff.gunadarma.ac.id

Julien Dubois
Université de Bourgogne
julien.dubois@ubourgogne.fr

Vincent Vajnovszki
Université de Bourgogne
vincent.vajnovszki@ubourgogne.fr

Asep Juarna
University of Gunadarma
ajuarna@staff.guandarma.ac.id

ABSTRACT
The present paper discusses the implementation of BWCA in lossless image compression. BWCA uses Burrows Wheeler Transform (BWT) as its main transform. As one of combinatorial compression algorithm which in particular reordered symbols according to their following context, it becomes one of promising approach in context modeling compression. BWT was initially created for text compression, and here we study the impact of BWCA method and its improvement when applied to image compression. Since this application is quite different from the original method aim, we analyze the pre- and post-processing influences of BWT.

Keywords
BWT, Lossless, image compression.

1. INTRODUCTION
Common image compression standard uses frequency transform such as Discrete Cosine Transform. We propose a completely different approach based on combinatorial transform, called Burrows Wheeler compression algorithm (BWCA). This approach has originally developed for text compression software such as BZIP2, but it has been recently applied to the image compression field [1, 2, 3, 4, 5, and 6]. The main transform of BWCA is Burrows Wheeler Transform (BWT). It is a context modeling compression that reordered symbols according to their following context, so its output contains many runs of repeated symbols. Since text compression is usually lossless, we implement BWCA in medical imaging that should be able to reconstruct every bit perfectly, thus lossless. BWCA results are compared with the existing compression standard such as JPEG and JPEG2000.

2. ORIGINAL METHOD OF BWCA
A typical Burrows Wheeler compression algorithm (BWCA) that has been proposed by Burrows and Wheeler for lossless text compression consists of 3 stages as seen in Figure 1, where [7]

- BWT is the Burrows Wheeler Transform itself, that tends to group similar characters together,
- GST is the Global Structure Transform, that change the next consecutive characters to zeros,
- EC is an Entropy Coding.

The stages are processed sequentially from left to right. The output of a previous stage becomes the input of the next stage. As stated before, the main transform of BWCA is BWT. This transform rearrange the input data using a sorting algorithm. The output contains are exactly the same with the input, one differing only in their ordering. Figure 2 gives a simple example how BWT works in a small image. Pixels are encoded by a pair of hexadecimal values. There are no repeated symbols if the image is scanned from left to right. BWT as a context based transform tends to group similar pixels together as seen in Figure 2(b). This transform does not reduce the data size; by contrast it adds a few bytes information as a primary index to decode the data.

Figure 1. Original scheme of BWCA.

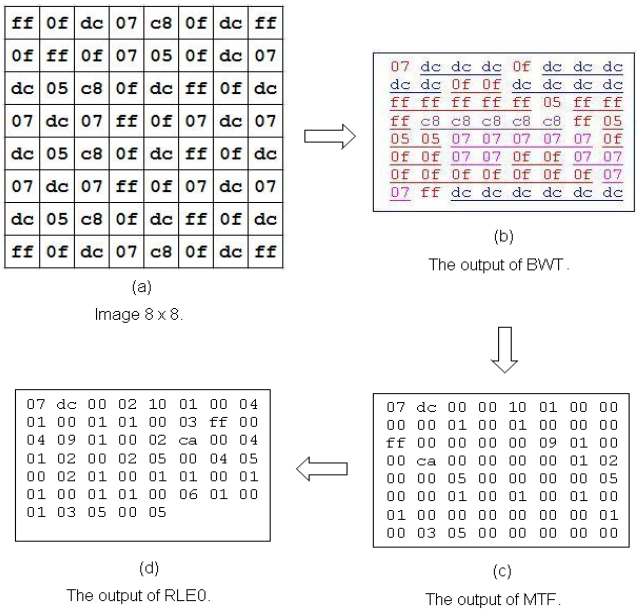


Figure 2. Example how BWCA works in a small image 8x8

The detail works how the BWT works and other transform will be explained below.

BWT

Figure 3 shows how the forward BWT works. In this example, we consider the first 16 bytes of the image array in Figure 2(a). BWT makes the rotations of input data as seen in Figure 3(a). Then, it sorts the rotations input pixel, see Figure 3(b). The last column, named L in Figure 3(b), of the obtain matrix together with the position of original data placed, are the output of BWT. Thus BWT output tends to groups similar pixels together. The primary index is equal to the position of the data in the original order.

1	ff	0f	dc	07	c8	0f	dc	ff	0f	ff	0f	07	05	0f	dc	07
2	0f	dc	07	c8	0f	dc	ff	0f	ff	0f	07	05	0f	dc	07	ff
3	dc	07	c8	0f	dc	ff	0f	ff	0f	07	05	0f	dc	07	ff	0f
4	07	c8	0f	dc	ff	0f	ff	0f	07	05	0f	dc	07	ff	0f	dc
5	c8	0f	dc	ff	0f	ff	0f	07	05	0f	dc	07	ff	0f	dc	07
6	0f	dc	ff	0f	ff	0f	07	05	0f	dc	07	ff	0f	dc	07	c8
7	dc	ff	0f	ff	0f	07	05	0f	dc	07	ff	0f	dc	07	c8	0f
8	ff	0f	ff	0f	07	05	0f	dc	07	ff	0f	dc	07	c8	0f	dc
9	0f	ff	0f	07	05	0f	dc	07	ff	0f	dc	07	c8	0f	dc	ff
10	ff	0f	07	05	0f	dc	07	ff	0f	dc	07	c8	0f	dc	ff	0f
11	0f	07	05	0f	dc	07	ff	0f	dc	07	c8	0f	dc	ff	0f	ff
12	07	05	0f	dc	07	ff	0f	dc	07	c8	0f	dc	ff	0f	ff	0f
13	05	0f	dc	07	ff	0f	dc	07	c8	0f	dc	ff	0f	ff	0f	07
14	0f	dc	07	ff	0f	dc	07	c8	0f	dc	ff	0f	ff	0f	07	05
15	dc	07	ff	0f	dc	07	c8	0f	dc	ff	0f	ff	0f	07	05	0f
16	07	ff	0f	dc	07	c8	0f	dc	ff	0f	ff	0f	07	05	0f	dc

(a)

	F														L
13	05	0f	dc	07	ff	0f	dc	07	c8	0f	dc	ff	0f	ff	0f	07
12	07	05	0f	dc	07	ff	0f	dc	07	c8	0f	dc	ff	0f	ff	0f
4	07	c8	0f	dc	ff	0f	ff	0f	07	05	0f	dc	07	ff	0f	dc
16	07	ff	0f	dc	07	c8	0f	dc	ff	0f	ff	0f	07	05	0f	dc
11	0f	07	05	0f	dc	07	ff	0f	dc	07	c8	0f	dc	ff	0f	ff
2	0f	dc	07	c8	0f	dc	ff	0f	ff	0f	07	05	0f	dc	07	ff
14	0f	dc	07	ff	0f	dc	07	c8	0f	dc	ff	0f	ff	0f	07	05
6	0f	dc	ff	0f	ff	0f	07	05	0f	dc	07	ff	0f	dc	07	c8
9	0f	ff	0f	07	05	0f	dc	07	ff	0f	dc	07	c8	0f	dc	ff
5	c8	0f	dc	ff	0f	ff	0f	07	05	0f	dc	07	ff	0f	dc	07
3	dc	07	c8	0f	dc	ff	0f	ff	0f	07	05	0f	dc	07	ff	0f
15	dc	07	ff	0f	dc	07	c8	0f	dc	ff	0f	ff	0f	07	05	0f
7	dc	ff	0f	ff	0f	07	05	0f	dc	07	ff	0f	dc	07	c8	0f
10	ff	0f	07	05	0f	dc	07	ff	0f	dc	07	c8	0f	dc	ff	0f
1	ff	0f	dc	07	c8	0f	dc	ff	0f	ff	0f	07	05	0f	dc	07
8	ff	0f	ff	0f	07	05	0f	dc	07	ff	0f	dc	07	c8	0f	dc

(b)

Figure 3. The BWT forward transform.

The reverse BWT principally is just another permutation of the original data [8]. Figure 4 shows how this process works. Started from position 15 where the original data is placed, refers to the context *ff* (as the first pixel of original data) and link 6, as a clue of next position to the second of original data. So, the next position is 6 refers to 0f and gives the next link to the next output. Therefore each step gives the permuted pixel value as output and will process the whole file, because of the cyclic rotations.

GST

Common GST that has been used by Burrows and Wheeler is Move-To-Front (MTF). As the second stage of original BWCA chain, MTF does not shrink data but it can help to reduce redundancy. MTF uses list update table as an index of MTF input. The list consists of 256 symbols since the input of grey level image are 256 pixel for 8 bit image per pixel. It processes the input symbols sequentially. Every input of MTF is moved to the front of the list, so the input symbols that occur often are transformed into small indices. The runs of repetitive symbols are transformed into zeros.

Position	input	context	link
1	07	05	7
2	0f	07	1
3	dc	07	10
4	dc	07	15
5	ff	0f	2
6	ff	0f	11
7	05	0f	12
8	c8	0f	13
9	ff	0f	14
10	07	c8	8
11	0f	dc	3
12	0f	dc	4
13	0f	dc	16
14	0f	ff	5
15	07	ff	6 ←
16	dc	ff	9

Figure 4. Inverse transform.

EC

There are two kinds of Entropy Coding that Burrows and Wheeler use in their original paper. First, they proposed to use Run Length Encoding Zeros (RLE0) after MTF, since there are a lot of zeros. Thus, RLE0 codes only the symbol zero to reduce the data size.

Finally to compress data efficiently, Burrows and Wheeler suggest implementing Huffman Coding or Arithmetic Coding (AC) to really compress the data [7]. Some approaches use Arithmetic Coding, which offers the best compression rates [9]. Further, Arithmetic Coding translates the entire data into numbers represented in certain base rather than translating each data symbol into a series of digit in certain base. Therefore AC approach is often more optimal than Huffman Coding.

3. CORPUS

The BWCA method has been created for lossless text compression and do not take into account the image's nature. Nevertheless, it can be applied successfully to images, as we will discuss in section 7. The lossless approach is obviously appropriate to medical image compression, which is expected to be lossless. Our experiments use 100 medical images from IRMA (Image Retrieval in Medical Applications) database [10] and Lukas Corpus [11]. The images are taken randomly in each category. IRMA database consists primary and secondary digitized X-ray films in portable network graphics (PNG) and tagged image file format (TIFF) format, 8 bits per pixel (8 bpp), examples of images are shown in Figure 5. The size of images is between 101 KB and 4684 KB. Lukas Corpus consists of 4 parts. We are using two dimensional 8 bit radiographs in TIF format.

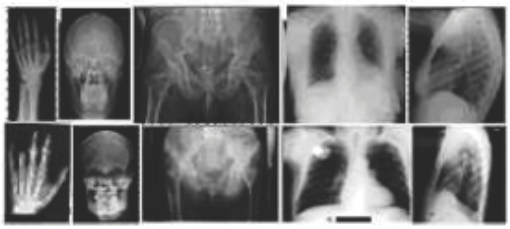


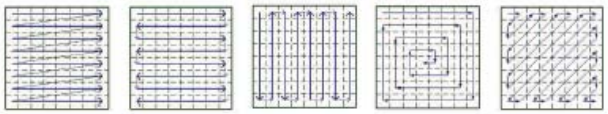
Figure 5. Example of tested images. From left to right : hand; head; pelvis; chest, frontal and lateral.

Table 1. Image compression ratios for different type of scan path.

image	L	LR	UD	Spiral	ZZ	8x8	8x8ZZ	3x3zz	Jpeg	J2K
Hand1	2.372	2.366	2.547	2.536	2.257	2.204	2.262	2.336	2.249	2.994
Hand2	2.260	2.251	2.390	2.382	2.091	2.052	2.073	2.138	2.205	2.769
Hand3	2.114	2.123	2.253	2.221	1.991	1.933	1.994	2.049	2.136	2.733
Hand4	2.685	2.679	2.830	2.802	2.551	2.411	2.510	2.557	2.189	2.909
Head1	2.219	2.216	2.274	2.273	2.155	2.108	2.174	2.220	1.992	2.554
Head2	2.481	2.480	2.527	2.538	2.366	2.337	2.392	2.466	2.210	2.938
Head3	2.566	2.565	2.527	2.563	2.350	2.303	2.349	2.422	2.363	2.932
Head4	2.726	2.721	2.721	2.764	2.544	2.502	2.542	2.622	2.399	2.548
Pelvis1	1.808	1.808	1.842	1.835	1.760	1.750	1.782	1.814	1.725	2.038
Pelvis2	1.850	1.848	1.890	1.876	1.806	1.791	1.829	1.863	1.797	2.105
AV. 10	2.308	2.306	2.380	2.379	2.187	2.139	2.190	2.249	2.109	2.665
Av. 100	2.516	2.515	2.577	2.575	2.357	2.315	2.364	2.442	2.280	2.924

4. LINEARIZATION SCHEME

BWCA is used to compress two-dimensional images, but the input of BWT is a one dimensional sequence. Thus, the image has to be converted from a two dimensional image into one dimensional sequence. This conversion is referred to as linearization or scan path. Some coding, such as Huffman Coding depend only on the frequency of occurrences of different gray values, therefore the scan path does not influence the compression performance, but another coding such as Arithmetic Coding and BWT itself, depend on the relative order of gray scale values and so are sensitive to the linearization method used.



(a) Left (L) (b) Left-right (LR) (c) Up-down (UD) (d) Spiral (S) (e) Zigzag (ZZ)

Figure 6. Linearization methods.

Some of the popular linearization schemes are given in Figure 6 [12]. We have tested 8 different linearization methods, scan image from left to right (L), left to right then right to left (LR), up to down then down to up (UD), zigzag (zz), spiral, divide image in small blocks 8x8, small blocks 8x8 in zagzag, and small blocks 3x3 in zigzag. Second column in Table 1 shows the compression ratios where the images is read conventionally (left right scanning). This result shows that BWCA original chain is better than JPEG but below JPEG2000. For the 10 tested images, only one image (Head4) gives better result than JPEG2000.

These preliminary results also show that BWCA results are better than JPEG but lower than JPEG2000. For more detail, 100 tested images, 91 provide better CR than JPEG, and among them 10 are better than JPEG2000.

5. BWT AND ITS IMPROVEMENT

BWT method is based on a sorting algorithm. There are several methods to improve the performance of sorting process, but they do not affect BWT results. Burrows and Wheeler suggested suffix tree to improve sorting process [7]. Other authors suggest suffix array or their own sorting algorithm [13, 14, and 15]. Figure 7 shows the relationship between BWT and suffix array for the same input in Figure 3, and we consider the input data is *Im*, then *n* is the length of *Im*, and so in this example $n = 16$.

The first and second column of Figure 7(a) is the given suffix for each of input data. The third column is a sorted suffix of the first column, and the fourth column is their suffix array (SA). The correlations between sorted rotations (the conventional of forward BWT) and sorted suffix are shown in the third column of Figure 7(b) and in the second column of Figure 7(c). They will be the same figure, if the adding symbols to create the rotations matrix of Figure 7(c) are ignored.

Suffixes	ID
ff 0f dc 07 c8 0f dc ff 0f ff 0f 07 05 0f dc 07	1
0f dc 07 c8 0f dc ff 0f ff 0f 07 05 0f dc 07	2
dc 07 c8 0f dc ff 0f ff 0f 07 05 0f dc 07	3
07 c8 0f dc ff 0f ff 0f 07 05 0f dc 07	4
c8 0f dc ff 0f ff 0f 07 05 0f dc 07	5
0f dc ff 0f ff 0f 07 05 0f dc 07	6
dc ff 0f ff 0f 07 05 0f dc 07	7
ff 0f ff 0f 07 05 0f dc 07	8
0f ff 0f 07 05 0f dc 07	9
ff 0f 07 05 0f dc 07	10
0f 07 05 0f dc 07	11
07 05 0f dc 07	12
05 0f dc 07	13
0f dc 07	14
dc 07	15
07	16

(a)

Sorted Suffixes	SA
05 0f dc 07	13
07 05 0f dc 07	12
07 c8 0f dc ff 0f ff 0f 07 05 0f dc 07	4
07	16
0f 07 05 0f dc 07	11
0f dc 07 c8 0f dc ff 0f ff 0f 07 05 0f dc 07	2
0f dc 07	14
0f dc ff 0f ff 0f 07 05 0f dc 07	6
0f ff 0f 07 05 0f dc 07	9
c8 0f dc ff 0f ff 0f 07 05 0f dc 07	5
dc 07 c8 0f dc ff 0f ff 0f 07 05 0f dc 07	3
dc 07	15
dc ff 0f ff 0f 07 05 0f dc 07	7
ff 0f 07 05 0f dc 07	10
ff 0f dc 07 c8 0f dc ff 0f ff 0f 07 05 0f dc 07	1
ff 0f ff 0f 07 05 0f dc 07	8

(b)

Post.	Sorted Rotations	L
13	05 0f dc 07 ff 0f dc 07 c8 0f dc ff 0f ff 0f 07	07
12	07 c8 0f dc ff 0f ff 0f 07 05 0f dc 07 ff 0f dc	dc
4	07 05 0f dc 07 ff 0f dc 07 c8 0f dc ff 0f ff 0f	0f
16	07 ff 0f dc 07 c8 0f dc ff 0f ff 0f 07 05 0f dc	dc
11	0f dc 07 c8 0f dc ff 0f ff 0f 07 05 0f dc 07 ff	ff
2	0f dc ff 0f ff 0f 07 05 0f dc 07 ff 0f dc 07 c8	c8
14	0f ff 0f 07 05 0f dc 07 ff 0f dc 07 c8 0f dc ff	ff
6	0f 07 05 0f dc 07 ff 0f dc 07 c8 0f dc ff 0f ff	ff
9	0f dc 07 ff 0f dc 07 c8 0f dc ff 0f ff 0f 07 05	05
5	c8 0f dc ff 0f ff 0f 07 05 0f dc 07 ff 0f dc 07	07
3	dc 07 c8 0f dc ff 0f ff 0f 07 05 0f dc 07 ff 0f	0f
15	dc ff 0f ff 0f 07 05 0f dc 07 ff 0f dc 07 c8 0f	0f
7	dc 07 ff 0f dc 07 c8 0f dc ff 0f ff 0f 07 05 0f	0f
10	ff 0f dc 07 c8 0f dc ff 0f ff 0f 07 05 0f dc 07	07
1	ff 0f ff 0f 07 05 0f dc 07 ff 0f dc 07 c8 0f dc	dc
8	ff 0f 07 05 0f dc 07 ff 0f dc 07 c8 0f dc ff 0f	0f

(c)

Figure 7. Relationship between BWT and suffix arrays.

For example the first column in Figure 7(b) was placed in line 13 of the original data, so the rotations matrix for this line starts from the symbol 13^{th} to 16^{th} then added the symbols 1^{st} to 12^{th} to complete the sorted rotations matrix. If the added symbols are omitted, this first line is the same with the first line of sorted suffixes in Figure 7(a). And the BWT output can be computed using suffix array (SA) as:

$$L(i) = \begin{cases} Im[SA[i]-1], & \text{if } SA[i] \neq 1 \\ Im[n], & \text{otherwise} \end{cases} \quad (1)$$

Hence, it does not need to create a sorted rotations matrix to obtain BWT output. Nevertheless, suffix sorting algorithms that run in linear time worst case is still open. Figure 8 shows different methods for BWT computation [16].

Method	complexity		
	time		space
	Worst-case	Avg.-case	Avg.-case
Ukkonen's suffix tree construction	O(n)	O(n)	-
McCreight's suffix tree construction	O(n)	O(n)	-
Kurtz-Balkenhol's suffix tree construction	O(n)	O(n)	10n
Farach's suffix tree construction	O(n log n)	O(n log n)	-
Manber-Myers's suffix array construction	O(n log n)	O(n log n)	8n
Sadakane's suffix array construction	O(n log n)	O(n log n)	9n
Larson-Sadakane's suffix array construction	O(n log n)	O(n log n)	8n
Itoh-Tanaka's suffix array construction	> O(n log n)	O(n log n)	5n
Burrows-Wheeler's sorting	O(n log n)	-	-
Bentley-Sedgewick's sorting	O(n log n)	O(n log n)	5n + stack
Sedward's sorting	O(n log n)	-	-

Figure 8. Different sorting algorithms used for BWT [16].

6. GST AND ITS MODIFICATIONS

As explain above, the main idea of MTF is to change runs of similar symbols into runs of zeros and also to locate frequent symbols near to the front of the list, therefore MTF output will be more convenient to compress by Entropy Coding. Most references show better results using MTF, especially for text compression. We have tested the BWCA original chain without MTF to see its effect, see the second column of Table 2. MTF increases BWCA performance till 4%. Even though 5 images give better results without MTF, but the average CR value for 10 and 100 images shows the improvement of CR. Balkenhol proposed the modification of MTF called M1FF [17]. The input symbol from the second position in the list update table is moved to the first position; meanwhile the input from higher positions is moved to the second position. Furthermore, Balkenhol gives a modification of M1FF called M1FF2. The symbol from the second position is moved to the first position of the list update table only when the previous transformed symbol was at the first position (M1FF2). Merely, the results of 100 tested images show that these transforms do not increase BWCA performance as shown in Table 2 column 3 and 4.

Albers has also presented other list update algorithms, called Time Stamp (TS). The deterministic version of this algorithm is TimeStamp(0) or TS(0). This method is also named a "Best 2 of 3" algorithm. It uses a double length list. Therefore, if 256 symbols are used in the input of BWT, the list for a "Best 2 of 3" will contain 512. So every symbol occurs twice. When a request is made, the position of an item is one plus the number of symbols for which both occurrences are in front of the second occurrence of the requested of symbol. Then the list is updated by moving that second occurrences to the front. Chapin improves the Ts(0) algorithm's and called it a "Best *x* of 2x-1" transform [18]. A "Bx" algorithm results in Table 2 are x=3 and 5.

Other variant of GST is Frequency Counting (FC). It based on the symbols ranking of their frequencies. It gives the highest rank to the symbol with the highest frequency. This transform is not very effective since it take time to favoring symbols [19]. Weighted Frequency Count (WFC) improves previous transform by defining a function based on symbol frequencies [16]. It also count the distance of occurrence symbol within a sliding window. The most occurrence symbol has a higher weight. This approach gives better compression ratios, but it is slower than FC because of the computation process [19]. Another GST method is Incremental Frequency Count (IFC) [9]. It is quite similar to WFC, but it is less complex but less performance than WFC.

Table 2. Comparative compression ratios for different GST variants using scan image up-down.

Image	no-MTF	MTF	M1FF	M1FF2	Ts(0)	Bx3	Bx5	FC	WFC	AWFC	IFC
Hand1	2.616	2.547	2.541	2.538	2.624	2.649	2.666	2.663	2.659	2.706	2.656
Hand2	2.196	2.390	2.387	2.387	2.449	2.466	2.476	2.470	2.484	2.513	2.475
Hand3	1.669	2.253	2.247	2.246	2.319	2.345	2.363	2.375	2.352	2.410	2.347
Hand4	2.589	2.830	2.825	2.824	2.920	2.939	2.946	2.938	2.979	3.000	2.954
Head1	2.251	2.274	2.263	2.262	2.329	2.349	2.364	2.357	2.348	2.400	2.359
Head2	2.561	2.527	2.518	2.518	2.590	2.613	2.632	2.631	2.611	2.672	2.614
Head3	2.554	2.527	2.52	2.519	2.606	2.632	2.650	2.646	2.645	2.695	2.635
Head4	2.751	2.721	2.714	2.714	2.801	2.824	2.841	2.832	2.857	2.902	2.839
Pelvis1	1.978	1.842	1.835	1.835	1.888	1.905	1.919	1.911	1.898	1.908	1.909
Pelvis2	2.026	1.890	1.881	1.880	1.936	1.952	1.966	1.961	1.951	1.950	1.960
Av.10	2.319	2.380	2.373	2.372	2.446	2.467	2.482	2.478	2.478	2.516	2.475
Av.100	2.475	2.577	2.570	2.570	2.654	2.679	2.696	2.694	2.694	2.724	2.687
St.Dev 10	0.351	0.324	0.325	0.325	0.338	0.340	0.340	0.339	0.353	0.364	0.342
St.Dev 100	0.575	0.584	0.586	0.586	0.607	0.640	0.614	0.613	0.623	0.618	0.614

Table 3. Comparative compression ratios for different GST variants using scan image spiral.

Image	no-MTF	MTF	M1FF	M1FF2	Ts(0)	Bx3	Bx5	FC	WFC	AWFC	IFC
Hand1	2.616	2.536	2.530	2.527	2.614	2.639	2.657	2.655	2.648	2.701	2.646
Hand2	2.193	2.382	2.380	2.380	2.453	2.472	2.486	2.479	2.481	2.517	2.476
Hand3	1.657	2.221	2.214	2.214	2.284	2.307	2.326	2.338	2.319	2.375	2.312
Hand4	2.587	2.802	2.800	2.798	2.898	2.923	2.937	2.926	2.954	2.985	2.927
Head1	2.251	2.273	2.263	2.263	2.329	2.350	2.365	2.358	2.347	2.399	2.359
Head2	2.576	2.538	2.528	2.528	2.602	2.626	2.646	2.645	2.623	2.689	2.627
Head3	2.578	2.563	2.555	2.553	2.642	2.669	2.689	2.685	2.684	2.736	2.674
Head4	2.784	2.764	2.757	2.756	2.848	2.874	2.894	2.885	2.905	2.957	2.887
Pelvis1	1.974	1.835	1.828	1.828	1.882	1.899	1.913	1.905	1.892	1.902	1.903
Pelvis2	2.018	1.876	1.867	1.867	1.922	1.939	1.954	1.948	1.937	1.932	1.946
Av.10	2.323	2.379	2.372	2.371	2.447	2.470	2.487	2.482	2.479	2.519	2.476
Av.100	2.477	2.575	2.568	2.567	2.651	2.677	2.694	2.692	2.691	2.721	2.685

7. EC AND ITS MODIFICATIONS

The original paper of Burrows and Wheeler uses RLE0 since there are a lot of zeroes after MTF [7]. The function of RLE is to support the probability estimation of the next stage. A long run of zeros tends to overestimate the global symbol probability. The previous best results use AWFC, up-down scan image method, RLE0 and AC. We omit RLE to see its impact on than chain. Table 4 shows the results of this test. RLE0 decrease BWCA compression ratios. Some authors separate the data stream and the runs so it does not interfere with the main data coding [8, 9], but it do not increase the performance of BWCA, as shown in Table 4 in the last two column. Here, we use modified RLE of [4], where all runs of size 2 or more are cut into 2 symbols and the length information is passed to a separate run length data stream and compressed separately by an Arithmetic Coding.

Table 4. Compression ratios results for RLE analysis.

Image	MTF			AWFC		
	RLE0	no-RLE	RLE2S	RLE0	no-RLE	RLE2S
Hand1	2.547	2.745	2.687	2.706	2.940	2.861
Hand2	2.390	2.489	2.479	2.513	2.611	2.600
Hand3	2.253	2.351	2.377	2.410	2.496	2.525
Hand4	2.830	2.973	2.953	3.000	3.156	3.120
Head1	2.274	2.370	2.360	2.400	2.490	2.480
Head2	2.527	2.668	2.650	2.672	2.816	2.793
Head3	2.527	2.703	2.664	2.695	2.890	2.837
Head4	2.721	2.920	2.873	2.902	3.128	3.058
Pelvis1	1.842	1.910	1.908	1.908	1.976	1.974
Pelvis2	1.890	1.973	1.967	1.950	2.034	2.027
Av. 10	2.380	2.510	2.492	2.516	2.654	2.628
Av. 100	2.577	2.738	2.706	2.724	2.890	2.849

8. CONCLUSION AND PERSPECTIFS

We presented the BWCA state of the art in lossless image compression. Each stage gives the important role to increase compression performance. Implement BWCA in image is not similar in text. We should consider pre-processing, where this stage could improve till 4% of compression performance. And based on our simulation, we propose to omit the RLE stage, because it decreases BWCA performance.

Our results show that BWCA is always better than JPEG but less than JPEG2000. From 100 image tested, 18% images give better CR than JPEG2000. But the standard deviation for those images is 0.525, meanwhile the standard deviation for 82% images which are less performance than JPEG2000 is 0.168. So, CR of BWCA compressed images that are better than JPEG2000 are much better than JPEG2000, meanwhile the others that are less performance, the CR differences are slightly small.

9. REFERENCES

[1] M. Ciavarella and A. Moffat, "Lossless image compression using pixel reordering," *Proceedings of* the twenty-seventh Australasian Computer Science Conference, 2004.

[2] N. R. Jalumuri, "A study of scanning paths for BWT based image compression," Master's thesis, 2004.

[3] X. Bai, J. S. Jin, and D. Feng, "Segmentation-based multilayer diagnosis lossless medical image compression," in VIP '05, pp. 9-14, Australian Computer Society, Inc., 2004.

[4] T. M. Lehmann, J. Abel, and C. Weis, "The Impact of lossless image compression to radiographs," vol. 6145, pp. 290-297, March 2006.

[5] Y. Wiseman, "Burrows-Wheeler based JPEG," *Data Science Journal*, vol. 6, pp. 19-27, 2007.

[6] E. Syahrul, J. Dubois, V. Vajnovszki, T. Saidani, and M. Atri, "Lossless image compression using Burrows Wheeler

Transform (methods and techniques)," in SITIS '08, pp. 338-343, 2008.

[7] M. Burrows and D. J. Wheeler, "A block-sorting lossless data compression algorithm," tech. rep., System Research Center (SRC) California, May 10 1994.

[8] P. M. Fenwick, "Burrows-Wheeler compression: principles and reflections," *Theor. Comput. Sci.*, vol. 387, no. 3, pp. 200-219, 2007.

[9] J. Abel, "Incremental frequency count-a post BWT-stage for the Burrows-Wheeler compression algorithm," *Softw. Pract. Exper.*, vol. 37, no. 3, pp. 247-265, 2007.

[10] T. M. Lehmann, M. O. GÄuld, C. Thies, B. Fischer, K. Spitzer, D. Keysers, H. Ney, M. Kohnen, H. Schubert, and B. B. Wein, "Content-based image retrieval in medical applications.," 2004.

[11] Lukas corpus. "http://www.data-compression.info/Corpora /Lukas Corpus/index.htm."

[12] S. Sahni, B. C. Vemuri, F. Chen, C. Kapoor, C. Leonard, and J. Fitzsimmons, "State of the art lossless image compression algorithms," 1997.

[13] S. Kurtz and B. Balkenhol, "Space efficient linear time computation of the Burrows and Wheeler-Transformation," in *complexity, Festschrift in* honors of Rudolf Ahlswede's 60th Birthday, pp. 375-384, 1999.

[14] G. Manzini, "Two spaces saving tricks for linear time LCP array computation," in *Proc. SWAT. Volume 3111* of Lecture Notes in Computer Science, pp. 372-383, Springer, 2004.

[15] P. Ferragina, R. Giancarlo, G. Manzini, and M. Sciortino, "Boosting textual compression in optimal linear time," *ACM*, vol. 52, pp. 688-713, July 2005.

[16] S. Deorowicz, "Universal lossless data compression algorithms. PhD thesis, Silesian University of Technology Faculty of Automatic Control, Electronics and Computer Science Institute of Computer Science, 2003.

[17] B. Balkenhol and Y. M. Shtarkov, "One attempt of a compression algorithm using the BWT," 1999.

[18] B. Chapin, "Switching between Two on-line list update algorithms for higher compression of Burrows-Wheeler transformed data," in Data Compression Conference, pp. 183-192, 2000.

[19] D. Adjeroh, T. Bell, and A. Mukherjee, The Burrows-Wheeler Transform: Data Compression, Suffix Arrays, and Pattern Matching. Springer US, June 2000.

Comparison of Random Gaussian and Partial Random Fourier Measurement in Compressive Sensing Using Iteratively Reweighted Least Squares Reconstruction

Endra
Department of Computer Engineering, University of Bina Nusantara
Jl K.H. Syahdan No.9 Kemanggisan / Palmerah, 11480, Jakarta, Indonesia
6221-53696930
endraoey@binus.edu

ABSTRACT

Compressive sensing is the recent technique of data acquisition where perfect reconstruction of signal can be made form far fewer samples or measurement than traditional Shannon-Nyquist sampling theorem. Iteratively reweighted least squares (IRLS) reconstruction is a compressive sensing reconstruction algorithm which is a first-order approximation to the *p*-norm minimization where $0 \leq p \leq 1$. In this paper, We compare the random Gaussian and partial random Fourier (using Discrete Cosine Transform) measurement to encode signal and then reconstruct the signal using IRLS algorithm for various *p*. From the numerical experiments, random Gaussian and partial random Fourier measurement, both give better perfect reconstruction probability for $p < 1$. Also both of them give almost the same perfect reconstruction probability as function of sparsity and measurement number, just slightly different for some of *p* value.

Keywords

Compressive sensing, IRLS, random Gaussian measurement, partial random Fourier measurement, perfect reconstruction probability, sparsity number, measurement number.

1. INTRODUCTION

Conventional approaches to sampling signal is using Shannon-Nyquist theorem : the sampling rate must be at least twice the maximum frequency in the signal (the so-called Nyquist rate) [1]. Compressive sensing (CS) is a sensing/sampling paradigm that goes against Shannon-Nyquist theorem. CS asserts that one can recover certain signals and images from far fewer samples or measurements than Shannon-Nyquist theorem use [2],[3]. Some potential applications are remote sensing [4], medical imaging [5], and sensor networks [6]. Three main issues in CS are sparsity of signal, CS measurement (Encoding) and CS reconstruction (Decoding). In this paper, signal will be considered sparse in time domain that contain a certain sparsity number which is number of non zero sample in signal. Random Gaussian and partial random Fourier matrices will be used to encode the signal. Both measurement will be compare by measuring the perfect reconstruction probability using Iteratively Reweighted Least Squares (IRLS) algorithms that was proposed in [7], [8].

2. IRLS ALGORITHMS

Consider an $M \times N$ measurement matrix Φ, where $M < N$, is used to encode signal x, result $y = \Phi x$, the vector of M measurements of an N dimensional signal x. One of widely known reconstruction algorithm is minimun ℓ_1 norm reconstruction :

$$\min_{x'} \|x'\|_1, \quad \text{subject to } \Phi x' = y \tag{1}$$

If measurement matrix, Φ, is random Gaussian distributed, there is a constant C such that if the sparsity of x has size K and $M \geq CK\log(N/K)$, then the solution to (1) will be exactly $x' = x$ [9], [10]. In [7] and [8] propose that ℓ_1 can be replaced with the ℓ_p norm, where $0 < p < 1$.

$$\min_{x'} \|x'\|_p^p, \quad \text{subject to } \Phi x' = y \tag{2}$$

In the case $p < 1$, IRLS can be used to for solving (2) by a replace the ℓ_p objective function in (2) by a weighted ℓ_2 norm [11] :

$$\min_{x'} \sum_{i=1}^{N} w_i x'^2_i, \quad \text{subject to } \Phi x' = y \tag{3}$$

The Eq. (2) can be written as :

$$\min_{x'} \sum_{i=1}^{N} \left| x'^{(n-1)}_i \right|^{p-2} x'^2_i, \quad \text{subject to } \Phi x' = y \tag{4}$$

where the weights, w, are computed from previous iterate $x'^{(n-1)}$, so that the objective in (3) is a first-order approximation to the ℓ_p objective : $w_i = \left| x'^{(n-1)}_i \right|^{p-2}$. The solution of (3) can be given explicitly, giving the next iterate $x'^{(n)}$ [8] :

$$x'^{(n)}_i = Q_n \Phi^T \left(\Phi Q_n \Phi^T \right)^{-1} y \tag{5}$$

where Q_n is the diagonal matrix with entries $1/w_i = \left|x'_i(n-1)\right|^{2-p}$. Using a small $\varepsilon > 0$ to regularize the optimization problem, w_i become :

$$w_i = \left|\left(x'_i(n-1)\right)^2 + \varepsilon\right|^{\frac{p}{2}-1} \quad (6)$$

3. NUMERICAL EXPERIMENTS

The Eq. (5) will be solved numerically using various parameters of sparsity number K , number of measurement M using random Gaussian that is set to be orthonormal and partial random Fourier matrices (in this paper using Discrete Cosinus Transfom (DCT) matrices) and p. The fix sample number N of signal x is 500 that has the sparsity number K which is choosed randomly with values 1 or -1 from a Gaussian distribution and using the sign function. Every experiment using certain parameters will be repeated 100 times to measure the perfect reconstruction probability. ε is initialized to 1 and $x'^{(0)}$ initialized to the minimum 2-norm solution of $\Phi x' = y$. The iteration (5) with w_i as in (6) is run until the change in relative 2-norm from the previous iterate is less than $\frac{\sqrt{\varepsilon}}{50}$, at which point ε is reduced by a factor 10, and the iteration repeated beginning with the previous solution. This process is continued through a minumum ε of 10^{-5}. The reconstruction is said to be perfect if mean squared error between x and x' less than 10^{-3}. In this paper, we are considering $0 \leq p \leq 1$.

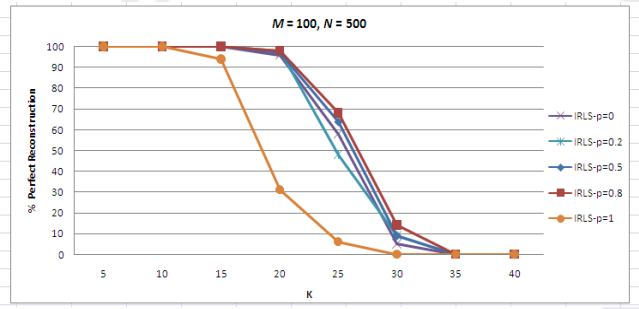


Figure 1. Perfect reconstruction probability as a function of *K* using random Gaussian measurement.

Figure. 1 shows the perfect reconstruction probability using random Gaussian measurement as a function of K for M = 100 (20 %), p = 0, 0.2, 0.5, 0.8, and 1. We can see for p < 1 give better result than p = 1, where for p < 1 perfect reconstruction can achieve 100 % until K = 20 while for p = 1 just until K = 10, after that decay more quickly than p < 1 and reach 0 % perfect reconstruction when K = 30. For p < 1 give almost the same results but the best is at p = 0.8, and for all reach 0 % perfect reconstruction when K = 35.

Fig. 2 shows the perfect reconstruction probability using random Gaussian measurement as a function of M in ratio of measurement numbers (RMN) = (M/N) *x 100 %* for K = 10, p = 0, 0.2, 0.5, 0.8, and 1. We can see again for p < 1 give better result than p = 1, where for p < 1 perfect reconstruction can achieve 100 % at MNR = 14 % while for p = 1 need until MNR = 18 %. For p < 1 give almost the same results but the best is again at p = 0.8, and for all reach 100 % perfect reconstruction when MNR = 18 %.

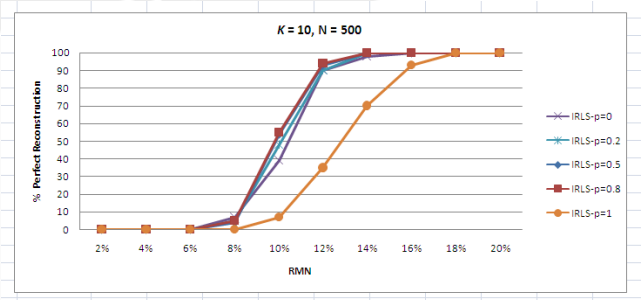


Figure 2. Perfect reconstruction probability as a function of *RMN* using random Gaussian measurement.

Figure. 3 shows the perfect reconstruction probability using partial random DCT measurement as a function of K for M = 100 (20 %), p = 0, 0.2, 0.5, 0.8, and 1.

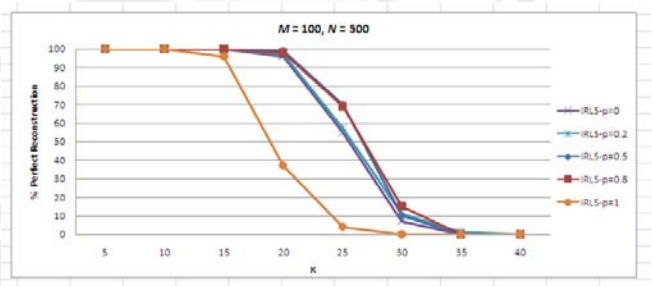


Figure 3. Perfect reconstruction probability as a function of *K* using partial random DCT measurement.

Figure. 4 shows the perfect reconstruction probability using partial random DCT measurement as a function of RMN for K = 10, p = 0, 0.2, 0.5, 0.8, and 1.

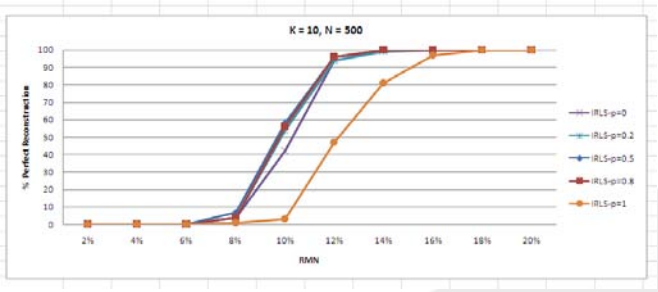


Figure 4. Perfect reconstruction probability as a function of *RMN* using partial random DCT measurement.

From Figure. 3 and 4, we can see that partial random DCT measurement give almost the same results as random Gaussian measurement. Although from Fig. 5, We can see that for $p = 0.2$ and 1, partial random DCT give slightly better perfect reconstruction probability as a function of K than random Gaussian measurement.

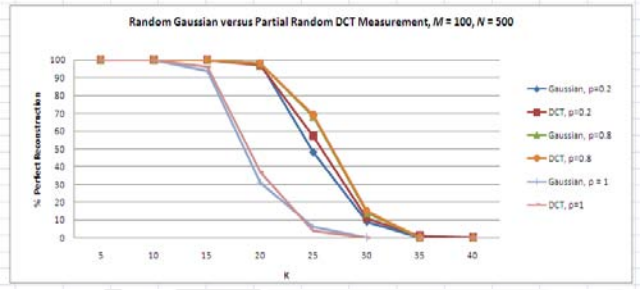


Figure 5. Comparison of perfect reconstruction probability as a function of *K* using random Gaussian and partial random DCT measurement.

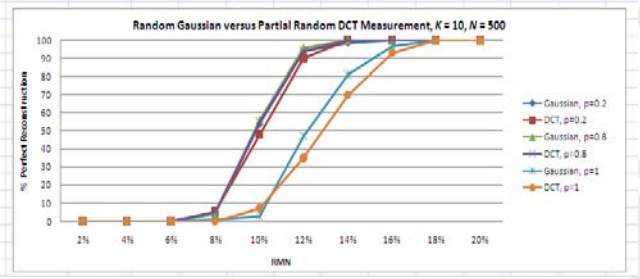


Figure 6. Comparison of perfect reconstruction probability as a function of *RMN* using random Gaussian and partial random DCT measurement.

From Figure. 6, We can see for perfect reconstruction probability as a function of *RMN*, the random Gaussian give better result than partial random DCT measurement for $p = 1$.

4. CONCLUSIONS

From the results above, We can conclude that random Gaussian and partial random Fourier (DCT) measurement, both give better perfect reconstruction probability for $p < 1$. Also both of them give almost the same perfect reconstruction probability as function of K and *RMN*, just slightly different for some of p value.

5. REFERENCES

[1] M. Unser. Sampling—50 Years after Shannon. *Proceedings of the IEEE*, 88(4):569–587, 2000.

[2] E.J. Candès and Michael B.Wakin, "An Introduction To Compressive Sampling," *IEEE Signal Processing Magazine*, March. 2008.

[3] D. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 52, no. 4, pp. 1289-1306, Apr. 2006.

[4] Jianwei Ma, "Single-Pixel Remote Sensing," *IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, VOL. 6, NO. 2, APRIL 2009*.

[5] M. Lustig, D.L. Donoho, and J.M. Pauly, "Rapid MR imaging with compressed sensing and randomly under-sampled 3DFT trajectories," in *Proc. 14th Ann.Meeting ISMRM*, Seattle, WA, May 2006.

[6] D. Baron, M.B. Wakin, M.F. Duarte, S. Sarvotham, and R.G. Baraniuk, "Distributed compressed sensing," 2005, Preprint.

[7] Chartrand, Rick., "Exact Reconstruction of Sparse Signals via Nonconvex Minimization," *IEEE Signal Processing Letters*, Vol 14, No. 10, Oct. 2007.

[8] Chartrand, R. and Yin. W, "Iteratively Reweighted Algorithms for Compressive Sensing", 2008, Preprint.

[9] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans.Inform. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[10] E. Candès and T. Tao, "Near optimal signal recovery from random projections: Universal encoding strategies?," *IEEE Trans. Inform. Theory*, vol. 52, no. 12, pp. 5406–5425, Dec. 2006.

[11] B. D. Rao and K. Kreutz-Delgado, "An affine scaling methodology for best basis selection," IEEE Trans. Signal Process., vol. 47, pp. 187–200, 1999.

Developing A Video Player Application for Phillips File Standard for Pictoral Data Format (NXPP): A Project View Approach

Eko Handoyo
Electrical Department Engineering Faculty
Diponegoro University
Jl. Prof. Sudharto, SH Tembalang
Semarang, Indonesia 50275
Telephone: +62247460057
eko.handoyo@undip.ac.id

Restiono Djati Kusumo
Information & Communication Technology
Fontys University of Applied Sciences
Hulsterweg 2-6 P.O. Box 141 5900 AC Venlo
Netherlands, the
Telephone: +31877879213
tion.kusumo@gmail.com

ABSTRACT
NXP is a leading semiconductor company who creates semiconductors that deliver better sensory experience in TVs, set-top boxes, identification applications, mobile phones, cars, and a wide range of other electronic devices. One of the groups in the company -Modem & Media Signal Processing group- is involved in developing and prototyping signal & video processing algorithms to be put in Systems-on-Silicon. The video data used here is in proprietary uncompressed format called PFSPD (Phillips File Standard for Pictoral Data). The assignment of this project is to develop a video player application capable to display output from PFSPD format video file. The application must have a high performance in real-time playback to the full limits of the hardware. The application is named "NXPP" (NXP PFSPD Player). This paper describes the architecture of the design that theoretically is able to provide the required performance. The method to determine the best design aspects was based on experiments to gather the performance data. It was realized by modifying design aspects in the development process supported with performance analysis. From the results of investigating the design aspects, the final design of NXPP includes the application of multithreading, Streaming SIMD Extension (SSE), OpenGL, Shader, and Direct Memory Access (DMA) technology.

Keywords
PFSPD, video player, multithreading, SSE, OpenGL, Shader, DMA.

1. INTRODUCTION
Systems-on-Silicon may contain digital, analog, mixed-signal, and often radio-frequency functions. The process of developing the algorithm includes testing the algorithm in a simulation, before finally it gets a proof-of-concept to be implemented in Systems-on-Silicon.

In Modem & Media Signal Processing group of NXP Semiconductors, The video data used is in proprietary uncompressed format called PFSPD (Phillips File Standard for Pictoral Data). Therefore, a support-tool capable of displaying output for this format is required in the development. With that tool, developers are able to compare the "original" and the "enhanced/processed" video. There are some existing applications used by the company which are considered having problems in performance, functionality, and maintenance.

The assignment for this project is to develop an upgraded video player application capable to display output from a PFSPD format video file. The playback application need to have high performance and also has to be well documented. This project is intended as a support for the infrastructure in Research sector, especially in Modem & Media Signal Processing group. The application will be named "NXPP" (NXP PFSPD Player).

2. RESEARCH ASSIGNMENT
2.1 Initial Condition
There are several existing video player tools used for playback purpose in the development process, namely PP (PFSPD Player), VideoSim, and HDD_Play. There are several problems with existing applications. PP is created around year 1996 and considered obsolete. PP has poor performance caused by bottle neck in data transfer to graphic card. Video playback in PP is done by loading data to the memory, and it is capable to playback in 120 Hz refresh rate, but limited to 200 image frames (on Windows 32-bit Operating System). HDD_Play is a hard-disk-streaming playback application developed by the company, but it is created specially to interface with a specific Video I/O Card (Bluefish). VideoSim meets most of all the requirements but lacks the documentation and there is no access to the source code since it is a software licensed by an external supplier. The company is looking for the possibility to develop an upgraded support tool for video player which is at least able to playback 120 Hz input video file, compatible with generic GPU, and capable of streaming playback (not only by loading data to the memory).

Short comparison between these video player applications is shown in table below:

Table 1. Comparison of existing application

	Playback from memory	Streaming playback from hard	Ease of use	Graphic Card API	Multi-files support	Support display hardware	*Source code access*
HDD_ Play		√	+	Blue fish		n.a.	√
PP	√		+	Dire ctX	√	+/-	√
VideoS		√	+	Dire		+	

im			/-	ctX			
NXPP (wish list)	√	√	+	*OpenGL*	√	+	√

PP's and HDD_Play's source code and documentation are available and provided by the company. The library required to access (read/write) PFSPD files is called CPFSPD. This library is a GPL (General Public License) open source library specially intended to handle PFSPD file format. There are also some experts in the company who can provide advice for me in the field of CFPSPD, Graphic Card Interface, etc.

2.2 Description of Assignment

In general, the strategies to accomplish the project are described as follows:

1. Research assignment

Research is done to study the requirements, existing applications, and available options for the application development.

2. Implementation

This is the main part of the project. It started with building a prototype with fundamental structure design of the application, to be followed with the next incremental design. The result of Research assignment includes basic knowledge about video processing and PFSPD file characteristics, basic requirements for initial development of NXPP, and initial application design to begin developing NXPP.

2.3 PFSPD format: YUV color space, sub-sampling, data structure

YUV is a color space for color image pipelining (series of data processing) where color image is encoded with taking human perception into account. In YUV, color will be defined by only 2 chrominance components (U and V) and Y component in YUV defines the luminance (brightness).

Lower sampling rate for UV means reduced chrominance components resolution. It is said that this is taking human perception into account because human eyes are more sensitive to luminance (brightness) resolution than chrominance (color details) resolution.

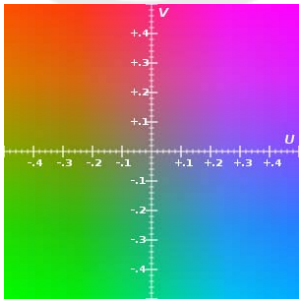


Figure 1. Example of U-V color plane with Y value = 0.5

As for the difference between YUV 4:2:2 and YUV 4:2:0 is shown below:

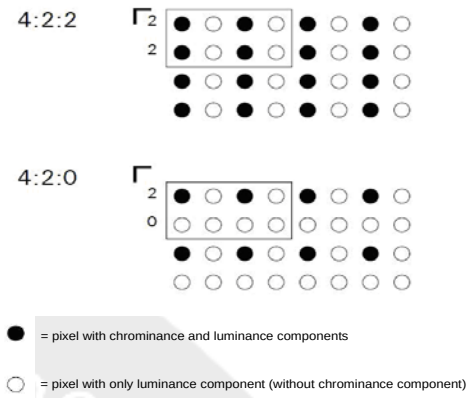


Figure 2. YUV 4:2:2 and 4:2:0 sub-sampling pattern

There are several kinds of data structures for PFSPD-YUV file. They are multiplexed, semi-planar, and planar.

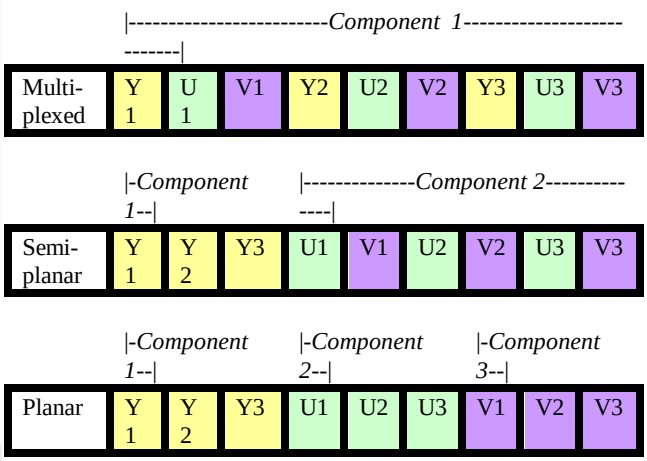


Figure 3. Multiplexed, semi-planar, and planar data structure example

For planar, each of Y-U-V values is stored separately. For multiplexed, the Y-U-V values are merged together. For semi-planar, Y component is separated but U & V components are merged.

2.4 Adopting HDD_Play Design

In Research assignment, it is decided to build NXPP from scratch by adopting HDD_Play design with multithreading architecture. In order to be able to do streaming playback of a video file, speed performance is important aspect in the design. That is why NXPP design is supported with multithreading, so the application is able to do different processing tasks in the same time.

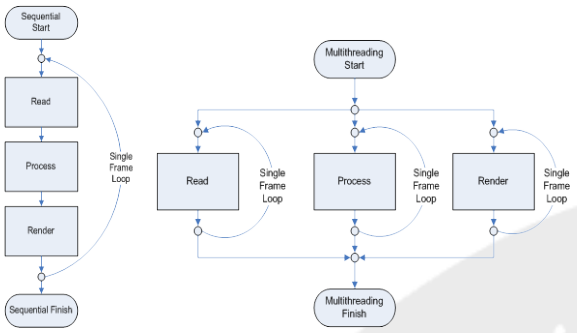


Figure 4. Example of sequential and multithreading flowcharts

3. IMPLEMENTATION

3.1 Building Application Structure

In this phase, NXPP is implemented by focusing on the application structure. NXPP structure is built with modular & multithreads architecture by taking example on HDD_Play design. In this phase, implementation is not focused on expanding the modules.

The idea of Phase 1 is to build the fundamental structure of NXPP application, so the next Phases will be to develop each module. In the end of this phase, NXPP application structure as fundamental design is obtained. The standalone version is built by using GLUT (GL Utility Toolkit). Once the standalone version was obtained, the next step would be integrating it with NXPP. However, GLUT does not seem compatible with multithreading management in NXPP.

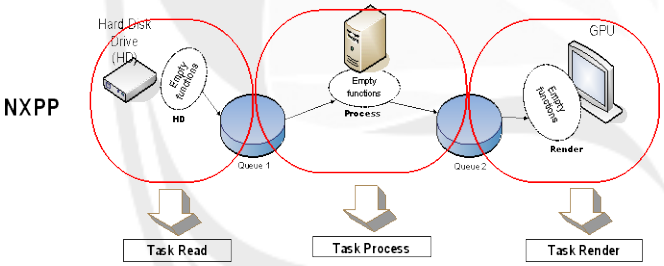


Figure 5. NXPP Implemented in Phase 1

3.2 Transfer Multiplexed Data to GPU

The objective in Phase 2 is to develop further the subsystems in NXPP application. As for the Render subsystem, it is decided that it will be developed by using OpenGL and Shader. To get acquainted with the Graphic API (OpenGL and Shader), a standalone version of the application is implemented before directly implementing the NXPP application design.

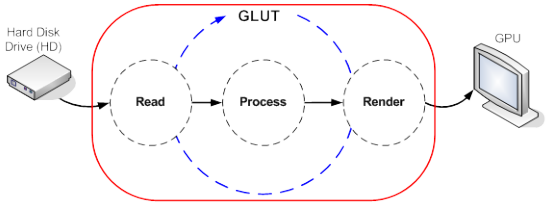


Figure 6. Concept of standalone version with single-threaded loop

Therefore, this Render subsystem is integrated with NXPP by using Freeglut instead of GLUT. The process of render (transferring data to graphic card) in this phase is done by transferring data in multiplexed format.

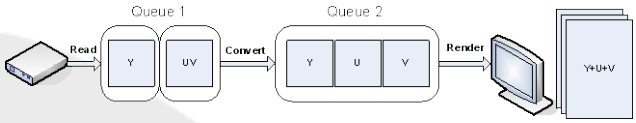


Figure 7. Concept of sending data in multiplexed format in NXPP

In the end of this phase, NXPP is able to playback video to the display by sending multiplexed data to the GPU. However, it doesn't seem to satisfy the requirement as it is showing a slow speed performance.

3.3 Transfer Semi-planar Data to GPU

The main objective in Phase 3 is to improve the speed performance of NXPP to fulfill the software requirement. To improve the speed performance a solution come up to reduce the bandwidth use for transferring data to GPU by sending data to GPU in semi-planar format instead of in multiplexed format.

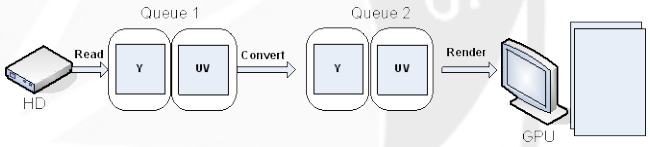


Figure 8. Concept of sending data to GPU in semi-planar format in NXPP

Testing on this version of NXPP shows that the speed performance for streaming playback a PFSPD file is close to the required speed performance requirement. However, there is some artifact/defect in the image shown in the display. Despite of improvement in the speed performance (around 100 frame per second compared to required 120 frame per second speed), the YUV color in the output image seems not constructed properly.

3.4 Transfer Planar Data to GPU

The objective in Phase 4 is to improve the speed performance of NXPP (to fulfill the software requirement) with correct output. A solution come up to solve this problem by rendering data in planar format to the GPU. Testing on this version of NXPP shows that the speed performance of the application nearly fulfils the speed performance in requirement. The applications manage to streaming playback PFSPD file in around 100 frames per second (the requirement is 120 frames per second) and with correct YUV image output.

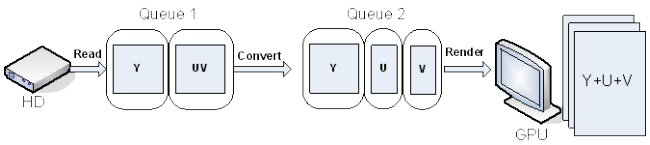


Figure 9. Concept of sending data to GPU in planar format in NXPP

3.5 Optimizing the Application

From phase 4, NXPP is able to do streaming playback with speed performance close to the speed performance required. It is believed that by rendering data in planar format is already using the bandwidth for data transfer to GPU effectively. However, the speed performance of the application is still not satisfying the requirement. The objective in this phase is to improve the speed performance of NXPP to satisfy the requirement, by optimizing the application.

SSE is implemented in the Convert module to improve the conversion process. SSE stands for Streaming SIMD Extension, an instruction set extension library created by Intel. SSE allows the processor to access data in 128-bits register, instead of the standard amount 32-bits (for 32-bit operating system). A register in a processor is a small amount of storage available in CPU, which allows quick access to it. With SSE, the CPU is expected to use less cycle for computation in the conversion process.

An idea also comes up to optimize NXPP design by avoiding Y value copy operation between queue 1 and queue 2.

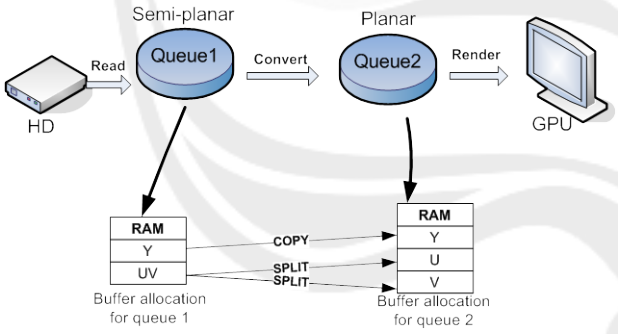


Figure 10. NXPP with 2-queues design

To optimizing the software design of NXPP, a universal buffer concept is implemented in NXPP system. The universal buffer is designed to be able to store both data in semi-planar format and planar format. With universal buffer concept, all queues in NXPP system are actually referring to a single pool of universal buffer. Diagram below explained the concept of universal buffer.

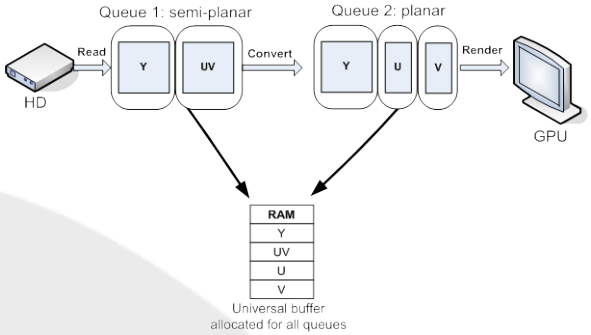


Figure 11. Universal buffer for all queues concept in NXPP

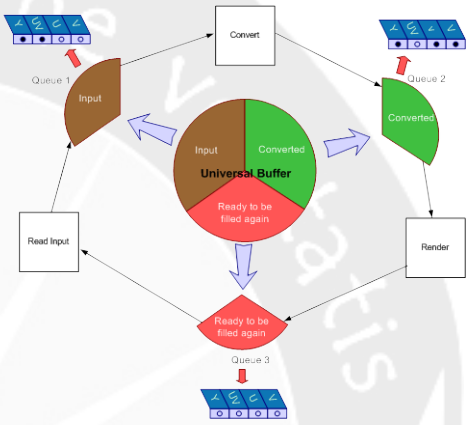


Figure 12. Universal buffer concept and 3 queues design in NXPP

The other optimization in NXPP is to implement Direct Memory Access (DMA) in rendering data to GPU. In the current version of NXPP, the process of rendering data to GPU is handled by CPU. It means that the CPU responsible for the data transfer from memory (RAM) to the GPU memory. With DMA, the data transfer from DMA-capable memory (RAM) to the GPU memory is handled by the GPU, which means reducing the CPU burden. Diagrams to compare the concepts of using DMA and without using DMA in NXPP is shown in figure 13 below.

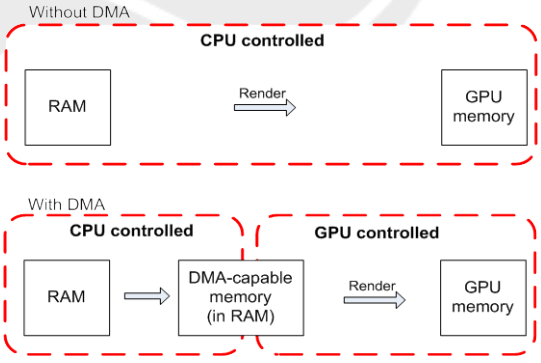


Figure 13. Comparison of without and with DMA implementation concept

DMA implementation in NXPP is done by using PBO (Pixel Buffer Object) from OpenGL. PBO is implemented in a way that

allows user to be able to switch between using and not using DMA (PBO). This ability to switch is able to be executed in run-time, so user can see the difference between using and not using DMA (PBO) instantly.

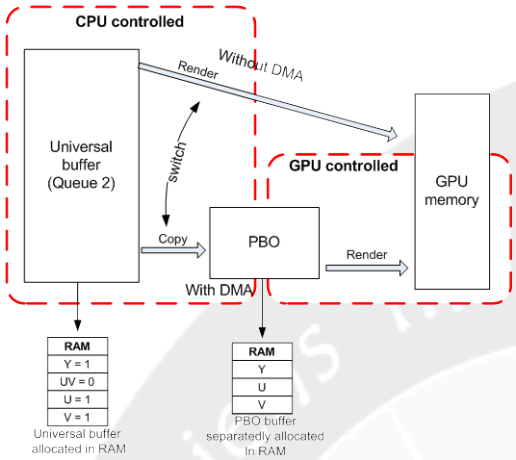


Figure 14. PBO implementation in NXPP

In the end of this phase, NXPP is designed with SSE implementation in the Conversion module, 3 queues with universal buffer concept, and with ability to use DMA in rendering data to GPU. The final design of NXPP is shown in figure below:

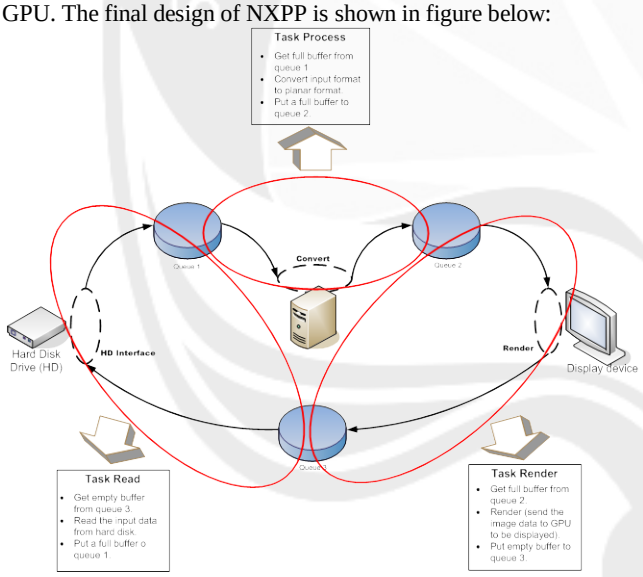


Figure 15. Final design of NXPP

Figure above shows the Tasks definition in NXPP: Task Read, Task Convert, and Task Render. This Tasks are executed in parallel processing with multithreading. Testing on the final version of NXPP shows that the speed performance is finally sufficient and fulfils the requirement.

4. CONCLUSION

The final version of NXPP is set as working prototype of NXPP for the company. It is called prototype since it only has limited features and still open to many possible improvements. However, the main challenge in NXPP which is to have the ability to streaming playback a Full-HD PFSPD file with 120Hz, is already achieved.
NXPP application is distributed among the colleagues in Modem & Media Signal Processing group of NXP, and first user trial gave positive responses. This application is a helpful contribution for the company to replace the obsolete software and licensed software currently in used. NXPP is used as one of the video player tool for the company, especially for Modem & Media Signal Processing group, and developed further for adding extra features or to cope with new requirements.
The latest version of NXPP (developed further) is even distributed to be used by NXP Semiconductors branches in San Jose (US) and Hamburg (Germany).

5. REFERENCES

[1] Ahn, Song Ho. “OpenGL Pixel Buffer Object (PBO)”. 2007. May 2009. <http://www.songho.ca/opengl/gl_pbo.html>.
[2] CPFSPD API description. 2008. May 2009. <http://pfspd.sourceforge.net/doxy_cpfspd/index.html>.
[3] Freeglut API documentation. 2003. June 2009. <http://freeglut.sourceforge.net/docs/api.php>.
[4] Goddeke, Dominik. “OpenGL Fast Transfers via PBO”. June 2009. <http://www.mathematik.uni-dortmund.de/~goeddeke/gpgpu/tutorial3.html#transfers>.
[5] “Integer Intrinsics Using Streaming SIMD Extensions (SSE) 2”. MSDN. June 2009. <http://msdn.microsoft.com/en-us/library/84t4h8ys.aspx>.
[6] Keith, Jack. Video Demystified. Fifth Edition. Oxford, UK: Elsevier, 2007.
[7] Kerr, Douglas A. Chrominance Sub-sampling in Digital Images. 2005. April 2009. <http://doug.kerr.home.att.net/pumpkin/Subsampling.pdf>.
[8] OpenGL Discussion and Help Forums. July 2009. <http://www.opengl.org/discussion_boards/>
[9] OpenGL Architecture Review Board. OpenGL Programming Guide, Sixth Edition. Addison-Wesley, 2007.
[10] Randima, Fernando., and Kilgard, Mark J. CG: The CG Tutorial. Addison- Wesley, 2003.
[11] “YUV”. Wikipedia. April 2009. <http://en.wikipedia.org/wiki/YUV>.
[12] “YUV to RGB conversion”. FOURCC. April 2009. <http://www.fourcc.org/fccyvrgb.php>.

Development Edge Detection Using Adhi Method, Case Study : Batik Sidomukti Motif

Adhi Pranoto
Informatics Engineering Department Atma Jaya University
adhi.pranoto@yahoo.com

Suyoto
Informatics Engineering Department Atma Jaya University
suyoto@mail.uajy.ac.id

ABSTRACT
This Paper shows new method for edge detection to analyze Batik motif. Adhi Method is robust method in edge detection which can detect accurate line for linear and unlinear line for many objects .The purpose of this research is develops new method in edge detection to analysis batik sidomukti motif which contains lot of oval and unlinear line. Adhi method is useful for edge detection batik sidomukti motif which is dominant oval and unlinear line.

Keywords
Batik motif, Edge Detection, Adhi Method, *Sidomukti*.

1. INTRODUCTION
Batik is one of Indoensian culture and listed by UNESCO as one of World Heritage in year 2009 reference number 170 [1][2]. Batik has wide motif as it demographic culture and every motif is uniqe[3]. Indonesia has wide variety of Batik motives due to plural culture and ethnics [3].
Edge detection used for identification of Batik Motif but none of method equal for detect the motif due to its complexity motif.
Purpose of this research is develops new edge detection method which is equal to edge detection for batik motif. This research focus to batik *Sidomukti* motif which has lot of oval object in it.
Most methods of edge detection methods are support for straight line and hard line[4][5][6] meanwhile Batik motif contain lot of soft line and flexible do the creator, especially handmade batik motif [7]. This far, none of the edge detection methods developed absolutely equal to batik motif due to batiks equity.
Outcome of this research is provides new edge detection method which is equal to batik motif and to support in data base science of batik identification as world heritage.

1.1 EDGE DETECTION
Edge detection is first step which cover information in image prcessing. Edge shows limtis of object and used for processing identification and segmenting object[8]. Purpose of edge detection is shows the line which surrounding the object [8].
Edge detectors are usually evaluated subjectively by observers and it is one of most commonly used operation in image analysis [4][6].Most of the objective evaluation methods assume knowledge of specific features such as known object boundaries in simple synthetic images. In such cases, the edge detection can be quantitatively measured, based on the known ideal detection considered to be the ground truth [5].

The problem of edge detection can be separated into three main subproblems [23]:

1. *Smoothing* : image intensities are smoothed via filtering or approximated by smooth analytic functions. The main motivations are to suppress noise and decompose edges at multiple scales.
2. *Differentiation*: amplifies the edges and creates more easily detectable simple geometric patterns.
3. *Decision*: edges are detected as peaks in the magnitude of the first-order derivatives or zero-crossings in the second-order derivatives, both compared with some threshold.

The Sobel edge detector [9] is one of the earliest edge detection methods. For many applications, it is used as a standard gradient computation method to retrieve the image gradient and edges [10].More specifically, the Sobel edge (1) detector contains two directional filters[9].

$$\text{Gx} = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, \text{Gy} = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} \quad (1)$$

The Canny edge detector [6] is one of the most widely used edge detectors in computer vision and image-processing community. In many applications, The Canny edge detector has been used as the standard image preprocessing technique [9].

$$\text{Gx} = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, \text{Gy} = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \quad (2)$$

Canny edge detection consists of four steps: noise suppression, gradient computation, non-maximal suppression, and hysteresis [6][10].
The Laplacian method of image highlights regions of rapid intensity change and is therefore often used for edge detection [29]. The Laplacian is often applied to an image that has first been smoothed with something approximating a Gaussian Smoothing filter in order to reduce its sensitivity to noise [30].

$$Gx = \begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix}, Gy = \begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix} \quad (3)$$

Edge detection has implement in wide area from small picture until complex picture. Edge detection methods usefull image processing for wide purpose such as medics, building, planning, and identification, recognition [11][12][13].
This research is development of earlier edge detection method using quick mask method (4) with matrix

$$Gx = \begin{bmatrix} -1 & 0 & -1 \\ 0 & 4 & 0 \\ -1 & 0 & -1 \end{bmatrix}, Gy = \begin{bmatrix} -1 & 0 & -1 \\ 0 & 4 & 0 \\ -1 & 0 & -1 \end{bmatrix} \quad (4)$$

Quick mask method has better result in edge detection than previous methods [11].
Quick Mask Methods can detect accurate edge line small picture [11]. Weakness of this method are unperfect color filtering and edge detection for high resolution picture. Multicolor image processing using this method will show the line in another color such as green, blue, red, especially the multicolor part lay in the center object [11]. Another weakness of its method is unoptimal edge detection for picture with lot of form in center of object. This method use positive value in matrix center and negative value in

the edge matrix [11]. It is good edge detection method for small size picture with color less.

There are five different criteria that are typically used for testing the quality of an edge detector[6]:

1. The probability of a false positive (marking something as an edge which isn't an edge)
2. The probability of a false negative (failing to mark an edge which actually exists)
3. The error in estimating the edge angle
4. The mean square distance of the edge estimate from the true edge
5. The algorithm's tolerance to distorted edges and features such as corners and junctions

Region-based technique uses the whole image to extract the information for the threshold value computation, while edge-based technique is based on the attributes along the contour between the object and the background [14]. Edge detection is more difficult than this for edges in textured regions and for colour pictures [24].

1.2 BATIK *SIDOMUKTI*

Batik is a famous and unique traditional heritage from Indonesia. Its uniqueness comes from its production process – which known as "mbatik", its motifs, and its values [10]. Since cultural and art product could become an economic product, as artistic and unique fabric product, batik could be a very valuable product economically, even in the modern era like today [15]. Indeed the point is the dominant design in batik[16]. Indonesia has wide variety of Batik motif and at least 23 provinces registered to UNESCO as sources of batik [17].

The items formed classical batik patterns or motifs can be broken down into two parts: the main and additional ornaments, and *isen-isen* (small ornaments used to fill the empty space in or between ornaments)[18][22]. The main ornaments define the motif style and they usually have some meaning. Some examples of the main ornament are *garuda*, *meru* (mountain), trees, temple, *parang*. Based on the ornaments and their structures, classical batik can be classified into two major classes, which are geometry and non-geometry[18][19].

Batik with natural motif is one of batik designs presenting natural descriptions such as animals, plants, fire, amulet, and the like[16]. The batik visually exposes reliefs like roots spreading to every direction. The relief is shown on *semen*, *sawat*, and *alas-alasan* motifs. The elements on the three motifs symbolize three groups of nature; the lower-level, the mid-level, and the upper-level nature[3][7].

Batik sidomukti is one of Yogyakarta's batik which is used for noble wedding ceremony [19]. The philosophy of *sidomukti* motif is honor and degree of the user [19]

1.3 ADHI METHOD

Different edge detection technique can be used for detecting the edge of a face image. For this specific case, Highpass filtering or edge detection technique was being applied on batik *sidomukti* motif. An edge is defined by a discontinuity in gray level values. In other words, an edge is the boundary between an object and the background [28]. Edge detection technique was observed better and is considered as the best for batik motif[20].

The term monochrome image or simply image refers to a two-dimensional light intensity function f(x, y), where x and y denote coordinates and the value of f at any point (x, y) is proportional to the brightness (or gray level) of the image at that point [21]. A digital image can be considered as a matrix whose row and column indices identify a point in the image and the corresponding matrix element value identifies the gray level at that point[21]. Value of matrix element correspondence to detection sensitivity, bigger value mean more sensitive in detection smooth object [26]. Edge points are characterized by a high local difference *(gradient)* in gray values [27].

Edge detection is applied after sharpening has been used for edge highlighting—for example, by gradient edge-generation pixels[24][26]. This operation may be performed by comparing the levels of adjacent pixels and, where the gradient is below some arbitrary level, reducing them to black Pixels with a gradient above that level are made white[24]. white line will be generated signifying a continuity of change corresponding to an edge on the image [24].

Adhi Method develops the quick mask method and it can cover the weakness of earlier method. Matrix of Adhi Method (5) is :

$$Gx = \begin{bmatrix} 0 & -3 & 0 \\ -3 & 12 & -3 \\ 0 & -3 & 0 \end{bmatrix}, Gy = \begin{bmatrix} 0 & -3 & 0 \\ -3 & 12 & -3 \\ 0 & -3 & 0 \end{bmatrix} \quad (5)$$

Adhi Method use bigger value in matrix element and the negative value lay in the straight line from center. Adhi Methods able to filter color better but case sensitive for red color. This method is quite excellent edge detection for artificial object and landscape object.

2. EXPERIMENTAL RESULT

Value of matrix element reached by several trial-error research to find the best performance. Ratio of positive value and negative value is 4:1 (abandon the negative mark) and the summary of negative value and positive value must be zero to show the best result. Positive value bigger than 12 will create noises by shows it color and noise from fiber canvas. Bigger value from this ration will increase the sensitivity result and shows the red color of object element.

For the edge-based thresholding technique, the idea of applying the boundary based attributes is based on the fact that discriminant features exist at the boundary between the object and the background[14].

Research follows by application program using visual C# 2006 for develops new edge detection method. User interface of program shows in figure 5.

Process of finding the best combination matrix value through several trial-error. Steps of experiment are:

1. Define the matrix size, matrix size we adopt from previous method, quick mask. Matrix size we used in this research are 3x3.
2. Matrix element position find by modification from quick mask method. Position of negative value and positive value influent the sensitivity of detecting edge object. Negative value play role in type of line which is detected, in this case we chose negative value on matrix element (1.2), (2.1), (2.3) and (3.2) for better result than previous method. Position of matrix element also correlated with object study which is geometry batik motif.
3. Define the matrix element ration between positive value and negative value. As our experience, ratio bigger than 4:1 for batik will shows noise due to it sensitivity, noise comes from canvas motif and red line color.
4. Define the matrix element value. Negative value and positive value. Due to step 3, matrix element we used are -3 and 12, that value 3 times to ratio. The matrix value find by several trial-error. Value of matrix element influencing the ability

and sensitivity of detecting edge object. Batik *sidomukti* has numbers of soft line and close each other.

5. Testing adhi method to several object study, we use several motif *sidomukti* to test its ability in edge detection. Testing with *sidomukti* motif shows the adhi method able to detect object.

Several trial-error in this experiment shows at table 1. Matrix combination shows the result in edge detection from various combinations. Value for matrix element -2 and 8, shows undetected the smooth object in the edge of picture and smooth objects in center of motif become one. Value for matrix element X -3 and 12 matrix element Y -2 and 8 vice versa shows that smooth object detected in the center motif and some smooth object in edge of picture undetected. Value for matrix component -3 and 12 show object in the center and edge picture detected .Value for matrix component more than -4 and 16 shows noise from canvas motif.

Value of matrix element influences the image processing. Negative value influences darkness object image processing and positive value influences brightness. Summary of matrix element which is not equal to zero will be dark or bright without show the edge line. Value of matrix element influences the ability in detecting smooth line and smooth objects. In case batik motif study combination of matrix element able to detect smooth object. bigger value increase ability of detecting object but increase it sensitivity to red color. Object which contain red color, event the object visually not shown red color, will be detected and shows in image processing as red line.

Adhi method through several trial-error experiments shows the ability in detecting smooth object. Smooth object are common object in *sidomukti* and other natural batik motif. Natural batik motif also contains red color due to its painting process and paint material. Red color detected in adhi method as red line. Object which is capture as red line in image processing shows object has red color component and shows due to its hue[23]. Hue is colour named from its subjective appearance and determined by the frequency of its radiated energy [24]. The sensation of a particular hue may be invoked either by the radiation of a particular wavelength (specified in nanometres) or by the wavelength generated by a mixture of colours [24].

Batik *Sidomukti* edge detection using Adhi Method as shown in figure 1 and figure 2 image processing using Adhi method give accurate edge detection result and it can detect the oval form in part of the object. We compare edge detection using different object as shows in figure 3 and figure 4. We chose object control with condition the object must contain extremely red color, rose red are object with extremely has red color. Edge detection of picture rose red shows red line in result of edge detection, the result shows the linear line and oval line detected. Red line detected in image processing due to its hue. This method shows that matrix case sensitive for red color. The difference between a colour and a specified reference colour of equal brightness, representing the hue and saturation values of that colour [24][25]

Adventage of Adhi method lay in the matrix form which is use diferrent value with the bigest value lay on the center. Positive value impact to brightness of image and we put big value to shows the un seen line of picture. Negative value impact in darkness, we spread the negative value in four part as image stabilizer and eliminate color component. Image processing shows the motif of batik *sidomukti*. Component of matrix must follow the rules :

a) Summary of positive value and negative value must equal to zero.
b) Positve value take place in the center of matrix
c) Negative value must quarter of positive value.
d) Location of negative value effected to type of line in edge detection
e) Red line in Adhi method comes from red color in source picture due to its hue.

Table 1. Result experiment ADHI method

Matrix X	Matrix Y	Result
$\begin{bmatrix}-1 & 0 & -1\\ 0 & 4 & -1\\ -1 & 0 & -1\end{bmatrix}$	$\begin{bmatrix}-1 & 0 & -1\\ 0 & 4 & -1\\ -1 & 0 & -1\end{bmatrix}$	Smooth object undetected in the center motif
$\begin{bmatrix}0 & -1 & 0\\ -1 & 4 & -1\\ 0 & -1 & 0\end{bmatrix}$	$\begin{bmatrix}0 & -1 & 0\\ -1 & 4 & -1\\ 0 & -1 & 0\end{bmatrix}$	Smooth object undetected in the center motif
$\begin{bmatrix}0 & -2 & 0\\ -2 & 8 & -2\\ 0 & -2 & 0\end{bmatrix}$	$\begin{bmatrix}0 & -2 & 0\\ -2 & 8 & -2\\ 0 & -2 & 0\end{bmatrix}$	Smooth objects detected become one object, smooth object in the edge of picture undetected
$\begin{bmatrix}0 & -3 & 0\\ -3 & 12 & -3\\ 0 & -3 & 0\end{bmatrix}$	$\begin{bmatrix}0 & -2 & 0\\ -2 & 8 & -2\\ 0 & -2 & 0\end{bmatrix}$	Smooth object detected in the center motif, some smooth object in the edge of picture undetected
$\begin{bmatrix}0 & -2 & 0\\ -2 & 8 & -2\\ 0 & -2 & 0\end{bmatrix}$	$\begin{bmatrix}0 & -3 & 0\\ -3 & 12 & -3\\ 0 & -3 & 0\end{bmatrix}$	Smooth object detected in the center motif, some smooth object in the edge of picture undetected
$\begin{bmatrix}0 & -3 & 0\\ -3 & 12 & -3\\ 0 & -3 & 0\end{bmatrix}$	$\begin{bmatrix}0 & -3 & 0\\ -3 & 12 & -3\\ 0 & -3 & 0\end{bmatrix}$	Smooth objects all detected
$\begin{bmatrix}0 & -4 & 0\\ -4 & 16 & -4\\ 0 & -4 & 0\end{bmatrix}$	$\begin{bmatrix}0 & -4 & 0\\ -4 & 16 & -4\\ 0 & -4 & 0\end{bmatrix}$	Noise from canvas motif

We also compare our method with previous edge detection methods. We compare with Sobel method, Canny method, Laplacian method and Quick mask method as shown in table 2. Sobel method and Canny methods have weakness in detecting unlinear lines, less sensitivity in detecting edge object, smooth lines and oval object. Lapacian method and Quick mask method have weakness in detecting smooth object int the center object.

Table 2. Comparison Previous Method

Method	Matrix	Result
Sobel	Matrix X : $\begin{vmatrix}-1 & 0 & 1\\ -2 & 0 & 2\\ -1 & 0 & 1\end{vmatrix}$ Matrix Y : $\begin{vmatrix}-1 & -2 & -1\\ 0 & 0 & 0\\ 1 & 2 & 1\end{vmatrix}$	Several smooth lines capture as single hard line. Cluster oval form detected as hard line. Small oval object undetected. Less precision on edge . Noise from canvas fiber in the center object
Canny	Matrix X : $\begin{vmatrix}-1 & 0 & 1\\ -2 & 0 & 2\\ -1 & 0 & 1\end{vmatrix}$	Several smooth lines capture as single hard line. Cluster oval form detected as hard line. Small oval object undetected.

	Matrix Y : $\begin{vmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{vmatrix}$	Less precision on edge . Noise from canvas fiber in the center object
Laplacian	$\begin{vmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{vmatrix}$	Smooth object undetected in the center motif Less sensitivity in the edge object.
Quick mask	$\begin{vmatrix} -1 & 0 & -1 \\ 0 & 4 & 0 \\ -1 & 0 & -1 \end{vmatrix}$	Smooth object undetected in the center motif Less sensitivity in the edge object

Figure 1. Batik *Sidomukti* Motif

Figure 2. Result of Adhi Method edge detection

Figure 3. Rose red

Figure 4. Result of Adhi Method edge detection

Figure 5. Edge detection program screen shoot

3. CONCLUSION

Adhi method as new edge detection method is support for identification edge line in batik *sidomukti* motif with limited noise. This method also compare to its previous method and shows noise reduction and higher performance. The result also compare with natural landscape object to show the accuracy and performance of this new edge detection method. Further research can be used for another batik motif to develop best formula which is equal for various batik motives.

4. REFERENCE

[1] Anonim, Convention for The Safeguarding of The Intangible Cultural Heritage, *UNESCO*, September 28-October 2, 2009

[2] Anonim, 2009, RI's Batik Named as World Heritage, The Jakarta Post, 6 august, 2009 [8]

[3] Hidajat, Robby., Study of Java Myth Structure-Symbol in Natural Element Batik Motif, *Bahasa dan Seni*, Tahun 32 No 2, Agustus, 2004 (in Bahasa)

[4] Tsibanov, V.N., A.M., Denisov., A.S.Kyrlov., Edge Detection Method by Thikanov Regularization, *International Conference Graphicon*, Moscow, 2004

[5] Yitzhaky, Yitzhak., Eli Peli., A Method for Objective Edge Detection Evaluation and Detector Parameter Selection, *IEEE*, Vol 25 No 8, August, 2003

[6] Nadernajad, Ehsan., Sara Sharifzadeh., Hamid Hassanpour., Edge Detection Technique: Evaluations and Comparisons, *Applied Mathematical Sciences*, Vol 2 No 31, 1507-1520, 2008

[7] Pujiyanto, 2003, Java Myth in Natural Element Batik Motif, *Bahasa dan Seni*, Tahun 31, No 1, Februari, 2003 (in Bahasa)

[8] Munir, Rinaldi, *Digital Image Processing*, Informatika, ITB-Press, Badung, 2004 (in Bahasa)

[9] Wang, Song., Feng Ge., Tiecheng Liu., Evaluating Edge Detection trhough Boundary Detection, *EURASIP*, Vol 2006 Article 76278:1-5, 2006

[10] Sabah, Mohammed., Jinan Fiadhi., Lei Yang., Morphological Analysis of Mammograms Using Visualization Pipelines, *Pak.J.Inform.and Technol*, Vol 2 No 2, 2003

[11] Sohail, Abu Sayeed MD., MD Rabiul Islam., Kaushik Roy., Real Time Recognigtion Using Neural Networks : A Secure Personnel Identification System, *ICECE 2004*,December 28-30, 2004

[12] Yong, Li., Wu Huayi., Adaptive Bulding Edge Detection by Combining LIDAR Data and Aerial Image, *Remote Sensing and Spatial Information Sciences*, Vol XXXVII, 2008

[13] Hoque,M.A., S.M.A.Razzak., M.A.K.Azad., A Simple Technique Applied to Edge Detection in Digital Image, *ICECE 2004*, December 28-30, 2004

[14] Mokji, M.M., S.A.R.Abubakar., Adaptive Tressholding Based on Co Occurrence Matrix Edge Information, *Journal of Computer*, Vol 2 No 8, October, 2007

[15] Wijaya,Mahendra., Multi Commercial Economy : The Development of Socio-Economic Network Complexity of Batik Industry in Surakarta, *CCSC*, Vol 5 No 8, August, 2009

[16] Khanafiah, Deni. and Hokky Situngkir, Computational Batik Motif Generation, Innovation of Traditional Heritage by Fractal Computation, ITB, unpublished, 2009

[17] Wacik, Jero., Commitment of Department of Culture and Tourism "Safeguarding of The Culture of Indonesian Batik", *Ministry of Culture and Tourism Republic Indonesia*, Jakarta, August 19, 2008

[18] Moertini, Veronica.S., Towards Classifying Classical Batik Images, *ICCT-UMB*, June 10, 2005

[19] Anonim, Catalog Yogyakarta's Batik, Kantor Wilayah Departemen Perindustrian Propinsi Daerah Istimewa Yogyakarta, Yogyakarta, pp 20, 1996

[20] Sidhi, Thomas Adi Purnomo, *New Edge Detection Method of Indonesian Batik Motif*, Atma Jaya Yogyakarta Univesity, Unpublish, 2009

[21] Anam, Sarawat., Md Sohidul Islam., M.A.Kashem., Real Time Face Recognition Using Step Error Tolerance BPN, *IACSIT*, Vol 1 No 1, April, 2009

[22] Anonim, Indonesia 2004, An Official Handbook, National Information Agency Republic of Indoensia, pp 22, 2004

[23] Bovik, Al,*The Essential Guide to Image Processing*, Academic Press, California, pp 186-320, 2009

[24] Cawkel, Tony, *Multimedia Handbook*, Routledge, London, pp 86-430, 1996

[25] Ebner, Marc, *Color Constancy*, John Wiley&Sons, West Sussex, England, pp 91, 2007

[26] Zhang, Yun, Texture-Integrated Classification of Urban Treed Areas in High-Resolution Color Infrared Imagery, *Photogrametric Engineering & Remote Sensing*, Vol 67 No 12, December, 2001

[27] Braunt, Thomas, Tutorial in Data Parellel Image Processing, *AJIIPS*, Vol 6 No 3, 2001

[28] Sharifi, Mohsen., Mahmoud Fatfhy., Maryam Tayefeh Mahnoudi, A Classified and Comparative Study of Edge Detection Algorims, *ITCC*, 2002

[29] Mostofizadeh, A., X.Shun., M.R.Kardan., Improvement of Nuclear Track Density Measurement Using Image Processing Techniques, *AJAS*, vol 5, 71-76, 2008

[30] Koschan, Andreas., Mongi Abidi., *Digital Color Image Processing*, John Wilwy & Sons, New Jersey, 2008

Discriminating Cystic and Non Cystic Mass Using GLCM and GLRLM-based Texture Features

Hari Wibawanto
Semarang State University
Faculty of Engineering
Sekaran, Gunungpati, Semarang
+62248508081
hariwibawanto@gmail.com

Adhi Susanto
Gadjahmada University
Faculty of Engineering
Jl. Grafika No. 2 Jogjakarta
+62274513665

Thomas Sri Widodo
Gadjahmada University
Faculty of Engineering
Jl. Grafika No. 2 Jogjakarta
+62274513665

S. Maesadji Tjokronegoro
Gadjahmada University
Faculty of Medicine
Jl. Farmasi, Jogjakarta

ABSTRACT
Research has been conducted to identify cystic mass and non-cystic mass in ultrasound images. A total of 127 images measuring 21x21 pixels, 82 images measuring 35x35 pixels, and 78 images measuring 50x50 pixels are taken as samples. Each image was transformed into a grey-level run-length matrix and a grey-level co-occurrence matrix. There were 11 features extracted from grey-level run-length matrix and eight features extracted from grey-level co-occurrence matrix, so that totally we have 19 features. The ability of features in distinguishing cystic mass and non-cystic mass images was determined by discriminant analysis, using statistical software package SPSS version 11.5. As a result, the 19 features extracted from grey-level run-length matrix and grey-level co-occurrence matrix could distinguishing cystic masses from non-cystic mass with an accuracy of 87.3% (for image size 21x21 pixels), 91.5 % (for image size 35x35 pixels), and 94.9% (for image size 50x50 pixels). Further analysis carried out by involving only 12 of the 19 features extracted, which consists of 5 features extracted from GLCM matrix and 7 features extracted from GLRL matrix. The 12 selected features are: Energy, Inertia, Entropy, Maxprob, Inverse, SRE, LRE, GLN, RLN, LGRE, HGRE, and SRLGE. Discriminant analysis with the 12 features as predictors can distinguish cystic mass image and non cystic mass with a level of accuracy of 85.3% (for image size 21x21 pixels), 91.5% (for image size 35x35 pixels), and 92.3% (for image size 50x50 pixels). Further analysis showing that Area Under the Receiver Operating Curve was 0.863 (for image size 21x21 pixels), 0.971 (for image size 35x35 pixels), and 0.995 (for image size 50x50 pixels), which means that the accuracy level of discrimination is good or very good. Based on that data, it concluded that texture analysis based on GLCM and GLRLM could distinguish cystic mass image and non-cystic mass image with considerably good result.

Keywords
Gray-level Co-Occurrence Matrix, Gray-level Run Length Matrix, ultrasound, cystic mass, non-cystic mass, texture features, textures analysis, discriminant analysis, Receiver Operating Characteristics.

1. INTRODUCTION
Ultrasonography (USG) is imaging techniques widely used in the diagnostic phase to capture image of the internal organs which can not be seen by human vision. The result of ultrasonography is black and white image representing an echo of the ultrasonic waves reflected by the layers of skin and internal organs. Such image can only be read and understood by medical or ultrasonography experts. Mistake in interpreting the ultrasound image can be dangerous, because the interpretation is becomes basis for further medical action. Human physical limitations, such as fatigue, may affect the results of interpretation of ultrasound images and in advance diagnostics mistaken could be happen.

Computer-aided image analysis has the potential capability to detect abnormal masses of tissue based on its ability to distinguish patterns of intensity variations in image generated by ultrasound. In 2-D shape, this variation is known as texture.

2. PROBLEMS STATEMENT
Texture analysis techniques will be applied to distinguish the cystic mass and non-cystic mass appearance in ultrasound images. Two techniques of features extraction, one is extraction based on grey-level co-occurrence matrix (GLCM) and the other is extraction based on grey-level run-length matrix (GLRLM) will be used to extract texture features. The problems are: Can the extracted features based on GLCM and GLRLM used to differentiate cystic masses and non-cystic mass in ultrasound image? What kind features that have appropriate contribution in distinguishing cystic mass image and cystic mass? How does the level of discrimination accuracy?

3. TEXTURE FEATURES EXTRACTION
3.1 Grey-Level Co-Occurrence Matrix
In a statistical texture analysis, texture features were computed on the basis of statistical distribution of pixel intensity at a given position relative to others in a matrix of pixel representing image. Depending on the number of pixels or dots in each combination, we have the first-order statistics, second-order statistics or higher-order statistics. Feature extraction based on grey-level co-occurrence matrix (GLCM) is the second-order statistics that can be use to analyzing image as a texture (Albregtsen, 2008). GLCM (also called gray tone spatial dependency matrix) is a tabulation of the frequencies or how often a combination of pixel brightness values in an image occurs.

The figure below represent the formation of the GLCM of the grey-level (4 level) image at the distance d = 1 and the direction of 0°.

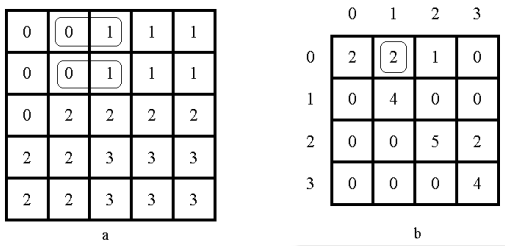


Figure 1 a. Example of an image with 4 grey level image b. GLCM for distance 1 and direction 0°

Figure 1.a. is an example matrix of pixels intensity representing image with 4 (four) levels of grey. Note the intensity level intensity level 0 and 1 are marked with a thin box. The thin box is representing pixel-intensity 0 and pixel intensity 1 as its neighbour (in the horizontal direction or the direction of 0 º). There are two occurrences of such pixels. Therefore, the GLCM matrix formed (Fig. 1.b.) with value 2 in row 0 and column 1. In the same way, GLCM matrix row-0 column 0 is also given a value of 2, because there are two occurrences in which pixels with value 0 has pixels 0 as its neighbour (horizontal direction). As a result, the pixels matrix representing in Figure 1.a. can be transformed into GLCM as Figure 1.b.

In addition to the horizontal direction (0 º), GLCM can also be formed for the direction of 45 º, 90 º and 135 º as shown in Figure 2 below.

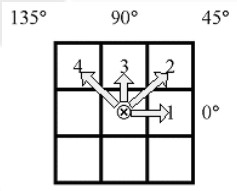


Figure 2. Direction of GLCM generation. From the center (⊗) to the pixel 1 representing direction = 0° with distance d =1, to the pixel 2 direction = 45° with distance d = 1, to the pixel 3 direction = 90° with distance d = 1, and to the pixel 4 direction = 135° with distance d = 1

Haralick and his colleagues (Haralick, Shanmugam, & Dinstein, 1973) extracting 14 features from the co-occurrence matrix, although in many applications only 8 (eight) features that are widely used, that is: Energy, Entropy, Max Probability, Inverse Diff. Moment, contrast, homogeneity, Inertia, and Correlation.

Although co-occurrence matrices capture the texture properties, it never directly used as a tool for analysis, such as comparing the two textures. The matrix of data must be extracted again to get the numbers that can be used to classify the texture. Haralick (Haralick, Shanmugam, & Dinstein, 1973) proposed 14 measures (or features), but Connors and Harlow in their study proposed, only 5 of 14 Haralick's features which are commonly used. These five features are: energy, entropy, correlation, homogeneity, and inertia (Kulak, 2002).

3.2 Gray-level Run-length Matrix

Grey-level run-length matrix (GLRLM) is a matrix from which the texture features can be extracted for texture analysis. Texture is understood as a pattern of grey intensity pixel in a particular direction from the reference pixels. Run length is the number of adjacent pixels that have the same grey intensity in a particular direction (Albregtsen, 1995). Gray-level run-length matrix is a two-dimensional matrix where each element $p(i, j \mid \theta)$ is the number of elements with the intensity i, in the direction θ. For example, Figure 3 below shows a matrix of size 4x4 pixel image with 4 gray levels. Figure 4 is a representation matrix GLRL (grey-level run-length) in the direction of 0º [$P(i, j \mid \theta = 0^o)$].

1	2	3	4
1	3	4	4
3	2	2	2
4	1	4	1

Figure 3. Matrix of image 4X4 pixels

Gray Level	Run Length (j)			
i	1	2	3	4
1	4	0	0	0
2	1	0	1	0
3	3	0	0	0
4	3	1	0	0

Figure 4. GLRL matrix

In addition to the 0º direction, GLRL matrix can also be formed in the other direction, i.e. 45º, 90º or 135º.

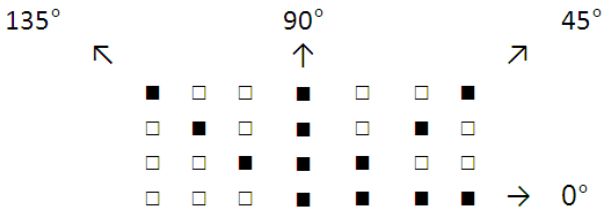


Figure 5. Run direction

Some texture features can be extracted from the GLRL matrix. Galloway (Tang, 1998) suggests 5 texture features based on this GLRL matrix, namely: Shot Runs Emphasis (SRE), Long Runs Emphasis (LRE), Gray Level Non-uniformity (GLN), Run Length Non-uniformity (RLN), and Run Percentage (RP).

Based on the observations that most of the features is only a function of $p_r(j)$, regardless of the grey level information contained in $p_g(i)$, Chu et al (Chu, Sehgal, & Greenleaf, 1990) adds 2 more features called Low Gray Level Run Emphasis (LGRE) and High Gray Level Run Emphasis (HGRE). This feature uses grey level of pixels in sequence and is intended to distinguish the texture that has the same value of SRE and LRE but have differences in the distribution of gray levels.

Dasarathy and Holder (1991) added 4 more features extracted from the matrix GLRL, namely: Short Run Low Gray-Level

Emphasis (SRLGE), Short Run High Gray Level Emphasis (SRHGE), Long Run Low Gray Level Emphasis (LRLGE), and Long Run High Gray Level Emphasis (LRHGE).

4. RESEARCH METHOD

The study began with the collection of ultrasound images in which there are cystic masses and non-cystic masses and has been proven true by ultrasound expert. Image taken directly from the video output of ultrasonograph, stored in a video cassette Hi-8, and then transferred into digital format using the video-processing software ULead Video Studio. The best frame containing the cystic and non-cystic masses, converted into 8-bit grayscale image. From that image, the region of interest (ROI) was cropped and analyzed.

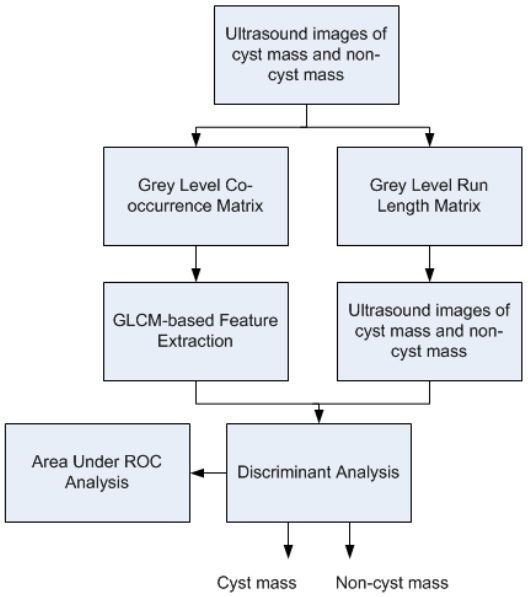


Figure 6. Steps in research

5. RESULTS AND DISCUSSIONS

A total of 127 images measuring 21x21 pixels consist of 38 cystic masses and 89 non-cystic masses, 82 images measuring 35x35 pixels consist of 30 cystic mass and 52 non-cystic masses, and 78 images measuring 50x50 pixels consist of 23 cystic masses and 55 non cystic masses, taken as samples. The results of discriminant analysis with SPSS version 11.5 statistical package software showed that the 19 texture features could differentiate cystic masses and non-cystic masses with an accuracy of 84.3% - 94.9%, depending on the size of samples and number of features using as predictors.

Table 1. Predicted accuracy of discriminant analysis for texture features based on GLCM and GLRLM

Size of Sample	GLCM-GLRL Texture Features	
	All Features	12 Features[1)]
21x21 pixel	87.3%	84.3%
35x35 pixel	91.5%	91.5%
50x50 pixel	94.9%	92.3%

Notes:

[1)] The 12 features are: SRE, LRE, GLN, RLN, LGRE, HGRE, SRLGE, Energi, Inertia, Entropi, Maxprob, Inverse

Performance evaluation accuracy of statistical prediction model (e.g. logistic regression or discriminant analysis) can also be done by ROC (receiver operating characteristics) curve analysis. ROC curve is a graphical plotting with the y-axis express sensitivity and the x-axis express false positive rate (Zou, O'Malley, & Mauri, 2007) (Ho, Goo, & Jo, 2004).

The following figure shows the ROC curve for discrimination using features based on GLCM and GLRLM as the predictors for the image size 21x21 pixels.

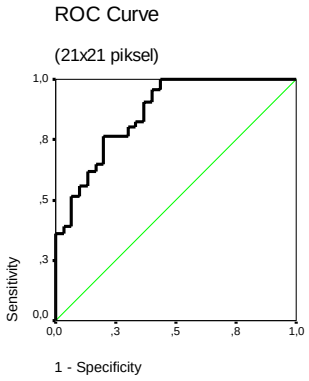


Figure 7. ROC curve for discrimination using GLCM and GLRLM features for image size 21x21 pixel

The level of accuracy is quantitatively expressed by the area under the ROC curve. Based on the ROC curve graphs generated by SPSS software version 11.5 shows that the area under curve is 0.863.

Table 2. AUROC (Area Under Receiver Operating Curve) of discrimination accuracy using glcm and glrlm features as predictors for image size 21x21 pixel

Area	Std. Error[(a)]	Asymptotic Sig.[(b)]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
.863	.039	.000	.786	.939

[(a)] Under the nonparametric assumption

[(b)] Null hypothesis: true area = 0.5

In general, the accuracy can be seen from the area under the ROC curve with the following general standards (Tape).

Table 3. Classifying level of accuracy based on area under ROC curve

Area Under ROC Curve	Classified as
0.90 – 1.00	Excellent
0.80 – 0.90	Good
0.70 – 0.80	Fair
0.60 – 0.70	Poor
0.50 – 0.60	Fail

Based on the above classification, the level of accuracy for the GLCM and GLRL-based texture features for image size 21x21 pixel classified as good.

The following figure shows the ROC curves for classification by using texture features based on GLCM and GLRLM as predictors for the image size 35x35 pixels.

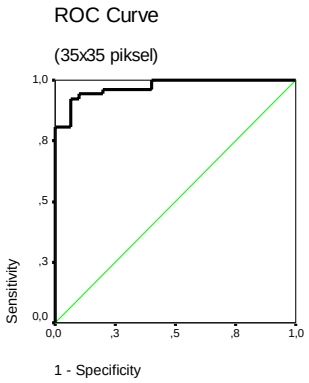


Figure 8. ROC curve for discrimination using GLCM and GLRLM features for image size 35x35 pixel

As seen in Figure 8, the level of accuracy is higher than the ROC curve for the image size of 21x21 pixels. Table 4 gives quantitative confirmation by showing the area under the ROC curve that is 0.971. Thus it can be stated that the level of prediction accuracy for the analysis of discrimination with features derived from the GLCM and GLRLM for image size 35x35 pixels is classified as very good.

Table 4. AUROC (*Area Under Receiver Operating Curve*) of discrimination accuracy using glcm and glrlm features as predictors for image size 35x35 pixel

Area	Std. Error[a]	Asymptotic Sig. [b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
.971	.015	.000	.942	1.001

[a] Under the nonparametric assumption

[b] Null hypothesis: true area = 0.5

The following figure shows the ROC curves for classification by using texture features based on GLCM and GLRLM as predictors for the image size 50x50 pixels.

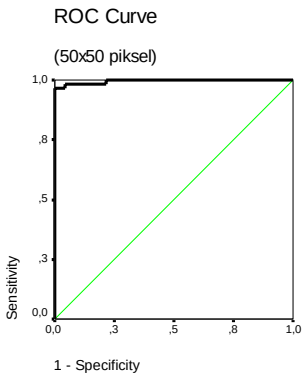


Figure 9. ROC Curve for discrimination using GLCM and GLRLM features for image Size 50x50 pixel

Figure 9 shows that its ROC curve is closer to the top y-axis, meaning that the level of prediction accuracy for discriminant analysis with texture features based on GLCM and GLRLM for 50x50 pixels image size is higher than the previous two ROC curves. Quantitative confirmation shown that the area under the curve generated by SPSS is 0.995, indicates that the prediction accuracy could be classified as very good.

Table 5. AUROC (*Area Under Receiver Operating Curve*) of discrimination accuracy using glcm and glrlm features as predictors for image size 50x50 pixel

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
.995	.005	.000	.986	1.005

[a] Under the nonparametric assumption

[b] Null hypothesis: true area = 0.5

Research by Khuzi and colleagues (Khuzi, Besar, & Wan Zaki, 2008) shows that the results is consistent with this research. In his research Khuzi using 4 features derived from GLCM, ie, contrast, energy, homogeneity, and correlation to detect the image of mammography with and without mass. GLCM features extracted in the four directions (0°, 45°, 90°, and 135°) with a region of interest size 8x8 pixels, 16x16 pixels, and 32x32 pixels. The findings of this study is that the best features to distinguish the image of the mammogram with mass and not mass are features based on GLCM with the ROI sized 8x8 pixels, in the direction of 0°.

In Karahaliou and fellow research (Karahaliou, et al., 2006), texture features based on LTEM (Laws Texture Energy Measure) for image size 128x128 pixels around microcalcification (MC) tissue was extracted. It conclude that features based on LTEM gives the best results for diagnosing the presence of breast cancer by 89% overall accuracy, sensitivity of 90.74% and specificity of 86.9%. These features are better than other features that are also extracted in this study, namely First Order Statistics (FOS), Gray Level Co-occurrence Matrices (GLCM), and Gray Level Run Length Matrices (GLRLM). For GLCM-based features, extraction was performed on the 4-direction of GLCM (0°, 45°, 90°, and 135°) with a distance d = 1. Thirteen features were derived from each GLCM. Specifically, the features studied are: Energy,

Entropy, Contrast, Local Homogeneity, Correlation, Shade, Promenance, Sum of Squares, Sum Average, Sum Entropy, Difference Entropy, Sum Variance and Difference Variance. Four values were obtained for each feature corresponding to the four matrices. The mean (M) and range (R) of these four values were calculated, comprising a total of twenty-six second order textural features. Twenty-six GLCM-based features that produce high predictive accuracy of 82%, higher than the predictions made by GLRLM based features which only reached 63% accuracy. These results are relevant to the study by authors, that is, texture features based on GLRLM has lower accuracy than texture features based on the GLCM in distinguishing cystic masses and non-cystic masses image.

Majumdar and Jayass (1999) in his research, classify seeds of wheat, barley, oats, and rye, using colour features and texture features as its predictors. Of the 25 textural features used in the discriminant analysis, 10 were GLCM features (mean, variance, uniformity, entropy, maximum probability, correlation, homogeneity, inertia, cluster shade, and cluster prominence), 12 were GLRM features (short-run, long-run, grey-level non-uniformity, run-length non-uniformity, run ratio, and GLRM entropy, and their ranges), and the remaining three were grey-level features (mean, variance, and range in grey-level). The research concluded that the variance, the GLCM based features, is the most important texture feature. Further stated that of the 10 most important texture features to classify images of grain, 5 (five) features derived from GLCM (entropy, correlation, mean, uniformity, and cluster Prominence) and 5 (five) features from GLRLM (variance, long run, short run, grey level non-uniformity, and run length non uniformity).

6. CONCLUSIONS

Based on research results and discussions, it conclude that:

a. GLRLM-based texture features can be used to distinguish between cystic masses and non-cystic masses on ultrasound images, with accuracy levels that are relatively lower than texture features based on GLCM and texture features based on combined GLRLM and GLCM.

b. GLCM-based texture features can be used to distinguish between cystic masses and non-cystic masses on ultrasound images, with accuracy levels higher than texture features based on GLRLM, but still lower than texture features based on combined GLRLM and GLCM.

c. Important texture features to distinguish cystic masses and non-cystic masses on ultrasound images are: SRE, LRE, GLN, RLN, LGRE, HGRE, SRLGE, Energy, Inertia, Entropy, Maxprob, Inverse.

7. ACKNOWLEDGMENTS

Our thanks to Dr. Sardjito General Hospital Yogyakarta for allowing us to use their valuable USG data.

8. REFERENCES

[1] Albregtsen, F. (2008, November 5). Statistical Texture Measures Computed from Gray Level Coocurrence Matrices. Retrieved April 30, 2010, from http://www.uio.no/studier/emner/matnat/ifi/INF4300/h08/undervisningsmateriale/glcm.pdf

[2] Albregtsen, F. (1995, November 14). Statistical Texture Measures Computed from Gray Level Run Length Matrices. Retrieved April 29, 2010, from http://www.ifi.uio.no/in384/info/glrm.ps

[3] Chu, A., Sehgal, C., & Greenleaf, J. (1990). Use of gray value distribution of run lengths for texture analysis. Pattern Recognition Letters , 415-420.

[4] Dasarathy, B., & Holder, E. (1991). Image Characterization Based on Joint Gray Level Run Length Distribution. Patt. Recog. Lett. , 497-502.

[5] Haralick, R. M., Shanmugam, K., & Dinstein, I. (1973). Texture Features for Image Classification. IEEE Transactions on Systems, Man, and Cybernetics , 610-621.

[6] Ho, P. S., Goo, J. M., & Jo, C.-H. (2004). Receiver Operating Characteristic (ROC) Curve: Practical Review for Radiologists. Korean Journal of Radiology (5), 11-18.

[7] Karahaliou, A. N., Boniatis, I. S., Skiadopoulos, S. G., Sakellaropoulos, F. N., Likaki, E., Panayiotakis, G. S., et al. (2006). A Texture Analysis Approach for Characterizing Microcalcifications on Mammograms., (p. The International Special Topic Conference on Information Technology in Biomedicine). Greece.

[8] Khuzi, A., Besar, R., & Wan Zaki, W. (2008). Texture Features Selection for Masses Detection in Digital Mammogram. 4th Kuala Lumpur International Conference on Biomedical Engineering. Kuala Lumpur: Springer.

[9] Kulak, E. (2002). Analysis of Textural Image Features for Content Based Retrieval. Thesis, Sabanci University.

[10] Majumdar, S., & Jayass, D. S. (1999). Classification of Bulk Samples of Cereal Grains using Machine Vision. J. Agric. Engng Res. (73), 35-47.

[11] Tang, X. (1998). Texture Information in Run-Length Matrices. IEEE Transactions on Image Processing , 7 (11), 1602-1609.

[12] Tape, T. G. (n.d.). The Area Under an ROC Curve. Retrieved January 7, 2009, from http://darwin.unmc.edu/dxtests/ROC3.htm

[13] Zou, K. H., O'Malley, J. A., & Mauri, L. (2007). Receiver-Operating Characteristic Analysis for Evaluating Diagnostic Tests and Predictive Models. Circulation: Journal of The American Heart Association

Fractal Terrain Generator

Budi Hartanto
University of Surabaya
Jl. Raya Kalirungkut
Surabaya
(031) 2981395
budi@ubaya.ac.id

Monica Widiasri
University of Surabaya
Jl. Raya Kalirungkut
Surabaya
(031) 2981395
monica@ubaya.ac.id

Gunawan Widjaja
University of Surabaya
Jl.Raya Kalirungkut
Surabaya
(031) 2981395

ABSTRACT
Simulated terrains have been extensively used in many flight simulators or games. Unfortunately, creating large terrains can be considered tedious and complex job. Therefore some flight game makers only created a limited amount of terrain and reused the same terrain for another level of the game. Though it seems solving the problem, increasing the number of terrains used in the program may make the game becomes more interesting. Since terrain actually is a fractal, the terrain can be generated using the concept of fractal. The resulted terrains can be saved in a text file to be utilized by any other programs. This research has overcome the terrain creation problem by generating user-defined terrain automatically.

Keywords
Terrain, Fractal, Flight Simulator, Flight Game.

1. INTRODUCTION
In many flight game programs, users have to fly an airplane through a sequence of certain terrain. The terrain usually undulates and it is one of the users' tasks to fly the plane through the hill and valley and to avoid collision with any of them.

In its simplest implementation, the game can create several hill and valley and store the data in a certain media. When the game needs to render the terrain, these stored hill and valley are read back from the media and rendered to the game environment. Due to the difficulty in hill and valley generation, there are only few hill and valley stored in the media. Because there are not many hill and valley can be taken from the media, the game will usually reuse the first available part of hill and valley and put them as the next hill and valley to be rendered on the game.

Though this approach can be used for several kind of game without loosing the users' interest, the availability of big supplies of hill and valley data can make the game more interesting and challenging. The problem is how can we provide affluent data of hill and valley easily?

One proposed solution to create affluent data of hill and valley is by generating them automatically. These hill and valley will build the terrain that can be used by other flight simulator or game program. A fractal method will be used to create the hill and valley in the terrain.

2. FRACTAL
Fractal is a pattern where partial or whole parts of the patterns have self similarity to the general pattern itself [4]. This fact will enable us to generate fractal recursively and heuristically. One example of fractal object is the Koch curve. The curve can be created by dividing a certain line into three parts. The middle part of the line then can be changed to a right triangle without its bottom part. If we recursively perform the same process to its sub-line, then we will get the Koch curve. Figure 1 shows several Koch curves created from a line, triangle, and polygon.

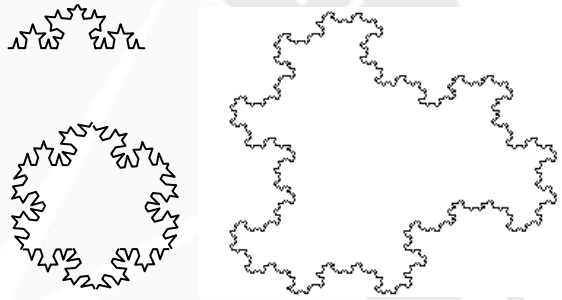


Figure 1. Several kinds of Koch curves.

If we magnify certain part of any Koch curves, we will get the similar pattern of the curve to its bigger part. This property is called self similarity and it becomes the major property of fractal object. As a matter of fact, many nature objects in our everyday life have this property. As an example a branch of a pine tree has a self similarity to the whole tree itself. A certain part of coast line has self similarity to its bigger part, etc. This phenomenon applies as well to terrain with its hills and valleys [6, 7]. A magnifier of a certain hill will give us several smaller hills and valleys as well.

Because of the terrain-self-similarity property it can be concluded that a terrain actually is a fractal in itself. Therefore it should be possible to generate terrain using any fractal method available.

3. GENERATING TERRAIN
3.1 Mid Point Displacement
Midpoint displacement is one of the fractal methods that can be used to create a terrain [1]. At the beginning set up four points that build up a square. Assign a certain height value to each point of the square and perform the following processes recursively:

- Divide the square to another four equal-size squares.
- Assign the height value at the middle of the big square. The value can be found by averaging the height value from all points in the big square corners. Add a modified

Gaussian-distribution-based random value to this average.

- Assign the height value at the middle point of each big square side. The value can be found by averaging the height value from its three adjacent points. Add a modified Gaussian-distribution-based random value to this average.
- Repeat the process to all newly generated squares until it reaches a certain set-up threshold..

The depth of the recursive process will determine the number of vertices that compose the grid. The relation of the number of vertices to the number of recursive level can be specified as:

Number of Vertices = $(2^{(n-1)}+1)^2$

Where *n* is equal to the number of recursive level. The number of vertices will determine the smoothness of the terrain. Increasing the number of recursive level will increase the number of vertices generated and as a consequence will increase the smoothness of the rendered terrain. However increasing the number of recursive level will increase the computation time and the memory space to store the on going processing data.

Data generated from the mid point displacement can be seen visually by simply connecting adjacent vertices with line. The row and column position of the squares can be used as the coordinate of the terrain and the height value of each point in the square can be used as the height of the terrain. In this simplest visualization, you can see the generated data of the terrain in a wireframe style visualization.

3.2 Gaussian Distribution

Gaussian distribution is commonly called normal distribution [8]. In this distribution, most data will cluster around the mean. That is why the curve of this distribution will appear as a bell. The function of the Gaussian distribution can be seen as follow:

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \quad \text{..... Eq. 1}$$

where:

μ = mean

σ = standar deviation

x = continuous random value at $-\infty < x < +\infty$

Figure 2 shows two Gaussian distributions. From the figure, it can be said that the mean in the Gaussian distribution *a* is less than the mean in the Gaussian distribution *b*. The mean of the Gaussian distribution is taken place at the peak of the shape. Meanwhile, from the overall shape of the shape, it can be said that the standard deviation of Gaussian distribution *a* is – again – less then the standard deviation of Gaussian distribution *b*. A Gaussian distribution with smaller standard deviation will have higher bell shape than the one with higher standard deviation.

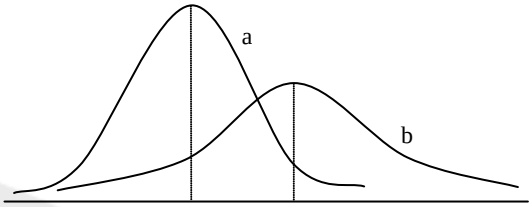


Figure 2. Gaussian bell-shape distribution

One property of the Gaussian distribution is:

$$\int_{-\infty}^{\infty} p(X)dx = 1 \quad \text{..... Eq. 2}$$

therefore:

$$\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx = 1 \quad \text{..... Eq. 3}$$

from this property it can be concluded that the value of μ and σ will always be in the range $0 \leq \mu, \sigma \leq 1$ because the total area of the curve is 1. To simplify the usage of Gaussian distribution, we need to convert the distribution to a standard Gaussian distribution. A standard Gaussian distribution is a Gaussian distribution with value 0 and 1 for the μ ,σ respectively [3, 8].

Random numbers that should be used in generating the terrain must confirm with this standard Gaussian distribution. Unfortunately many computer programming languange does not provide standard Gaussian distribution. Instead, most computer programming language provide a uniformly distribution random number.

In order to solve this problem, Goerge Edward Box and Mervin Muller introduced a Box-Muller transform that can be used to translate uniformly distributed random numbers to standard Gaussian distribution random numbers [9, 10]. The method takes two uniformly distributed random numbers and maps them to two independent standard Gaussian distributed random numbers.

Figure 3 shows the illustration of random numbers mapping from the uniform to Gaussian. In Box-Muller transform, two uniformly random number x_1 and x_2 which are distributed between 0 and 1 will map to two Gaussian numbers z_1 and z_2 by the following functions:

$$z_1 = \sqrt{-2\ln x_1}.\cos(2\pi x_2) \quad \text{..... Eq.4}$$

$$z_2 = \sqrt{-2\ln x_1}.\sin(2\pi x_2) \quad \text{..... Eq.5}$$

Unfortunately, calculating sine and cosine for many repeated processes can be considered costly. Therefore it is necessary to modify the functions and solve the equation in polar form. Given x_1 and x_2 in the range -1 to 1, set the value of *s* as $x_1^2 + x_2^2$. If $s=0$ or $s \geq 1$ discard x_1 and x_2 and find another x_1 and x_2 pair that generate s in the open interval 0 to 1.

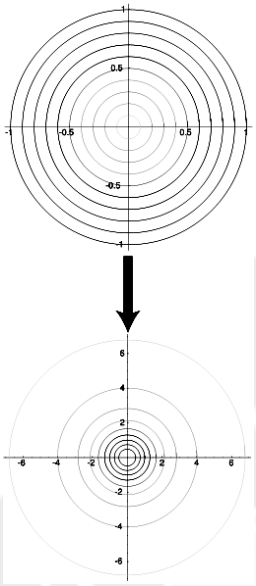


Figure 3. Uniform to Gaussian distributed random numbers

In the polar form, the mapping of uniform distribution to Gaussian can be expressed as [5]:

$$z_1 = x_1 . \frac{\sqrt{-2\ln s}}{s} \quad \text{..... Eq.6}$$

and

$$z_2 = x_2 . \frac{\sqrt{-2\ln s}}{s} \quad \text{..... Eq.7}$$

3.3 Modified Gaussian Random Value

Before the Gaussian distribution random value can be added to the height field, modify this value to include the recursive level into account. The recursive level determines the number of hills and valleys generated by the program.

A scaling factor will be applied to the Gaussian-distribution-based random value. The scaling factor that will be used is specified as:

$$d_n = (0.5)^{n.H/2} d \quad \text{..... Eq.8}$$

where:

d_n = scaling factor at recursive level n

n = level of the recursive process

H = roughness level of the surface

d = initial scaling factor

From the formula, it can be concluded that as the recursive level is getting higher, the scaling factor is getting smaller. This behavior matches perfectly with the natural terrain property. Most points in the terrain will have big height differences at the lower part of the hill and small height differences at the pinnacle.

Meanwhile the parameter H in the formula can be used to alter the roughness of a surface. Smaller value will create a rougher surface, and greater value will create smoother surface. Figure 4 show the difference of result using different values of H.

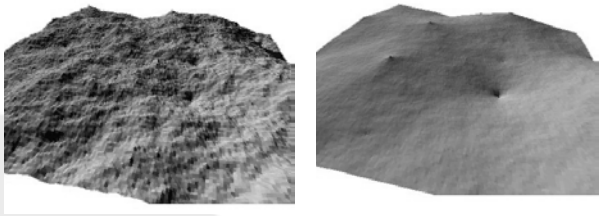


Figure 4. Surface roughness as determined by the value of H

3.4 Building and Storing the Terrain

In order to be able to be used by any other program, the generated terrain data must be stored in a file with general format that can easily be opened and read by any other program. In this research, the file format that is used for storing the terrain is a text file.

Three dimensional objects can be built from a collection of mesh triangles. The triangle shape is chosen because triangle is the simplest convex object. Therefore this object can not be divided anymore to another object especially to a non convex object. Guaranteeing convexity of an object is compulsory in object‘s rendering process.

Since the result of mid point displacement will generate mesh squares (see figure 5) then it is necessary to change the mesh squares to mesh triangles. Figure 5 shows the modification that must be made to change the generated mesh squares to mesh triangles.

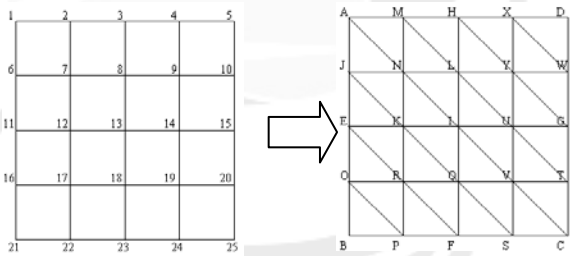


Figure 5. Surface roughness as determined by the value of H

One final process that must be performed after having all the required data is definitely storing the data itself. Two kinds of data that is compulsory to be stored are the vertices and faces. Figure 5 actually have already given clues about the vertices and faces relation to the indices of data. Numbers at the left picture of figure 5 shows the indices used for the vertices, while letters at the right picture of figure 5 shows the relation of indices to the faces of the mesh triangles.

Therefore, if figure 5 represents a certain generated terrain then there will be 25 indices that will be used to store all vertices. Meanwhile the number of faces to be generated will follow the following formula:

$$faces = 2.(\sqrt{vertices} - 1)^2 \quad \text{..... Eq.9}$$

In this case, the generated terrain in figure 5 will have 32 faces.

To avoid lighting problem in rendering process, all faces must be specified in either clockwise or counter clockwise manner. In this program, all faces will be stored in counter clockwise manner. Therefore the first and the second faces for terrain such as shown in figure 5 will be AJN and ANM respectively.

In general, data that will be stored in the text file are:

- The number of vertices and faces
- All vertices that build up the terrain. Each vertex will contain its three dimensional position (x, y, z)
- All faces that build up the terrain. Each face will refer to vertices index that build the face. Therefore instead of storing A, J, N, as the first face of terrain in figure 5, the application will store 1, 6, 7 as face's vertex indices.

3.5 Rendering

In order to prove the correctness of the terrain data generated by the program, a rendering facility will be built to show the terrain data visually. An openGL graphics library [2] will be used in the program to assist the process of terrain visualization. By this library, one can set up a camera-like viewing to see the terrain three dimensionally.

To increase realism of the terrain, a certain color scheme will be set up and used for coloring the terrain. Several certain heights will be used as a threshold for certain colors. Therefore, hill will have different color from valley. A texture mapping is used as well in the program to improve the realism of the terrain.

4. RESULT AND DISCUSSION

Figure 6 to 9 show several terrains that have been built using the application. Several parameters that influence the shape of the generated terrain are:

- The maximum number of recursive level (Iteration)
- The number of squares used (Tile)
- The roughness of terrain (Rough)
- Initial scaling factor (Scale)

Figure 6, shows terrain with many low hills. This kind of terrain can be generated using lower value of scale. In this example, the value of the scale is 50. On the other hand, increasing the scale value will yield terrain with many high hills. Figure 7 shows this kind of terrain. In this example, 75 is chosen as the value of the scale.

The roughness of the terrain can be set up as well in the program. Figure 6 shows the roughness of the terrain with value 3, while figure 7 shows the roughness of the terrain with value 2. As explained in sub chapter 3.3., smaller value of this parameter will make the terrain rougher.

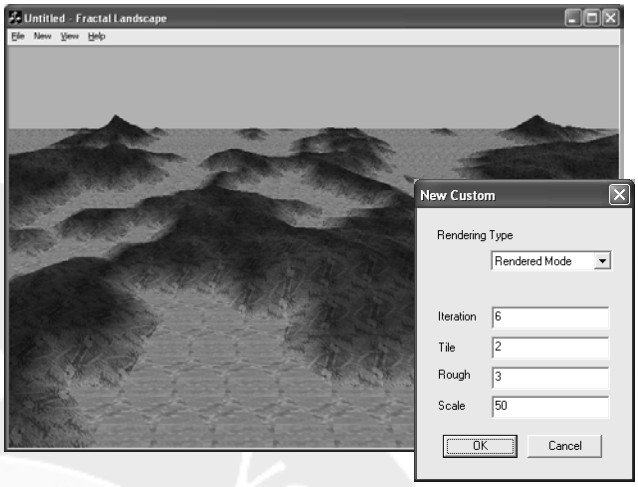


Figure 6. Lowland terrain

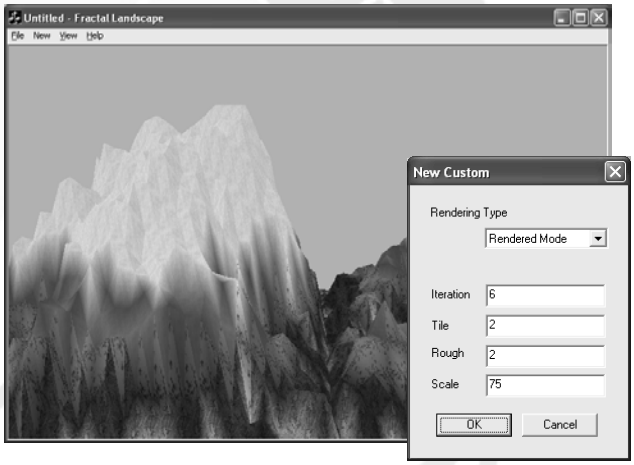


Figure 7. Highland terrain

With the assistance from openGL graphics library, one can view the terrain with a camera-like viewing. Therefore we can wander the terrain such as a flying bird. Figure 8 actually shows the same terrain such as shown in figure 7, after we fly closer to the terrain.

Two last figures, i.e. figure 9 and figure 10, show the genuine data of the terrain. Figure 9 shows the data in visualized meshed triangle, while figure 10 shows the raw data in text file. Each vertex and face data in the text file is always preceded with the keyword "v" and "f". This keyword will ease other programmers to read and interpret the data for their own use.

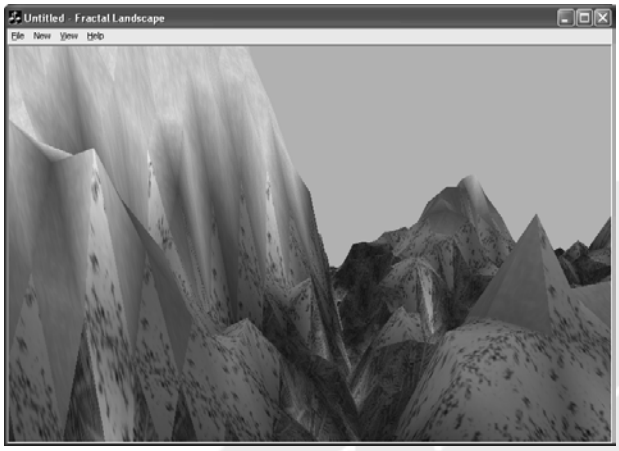


Figure 8. Walkthrough the terrain

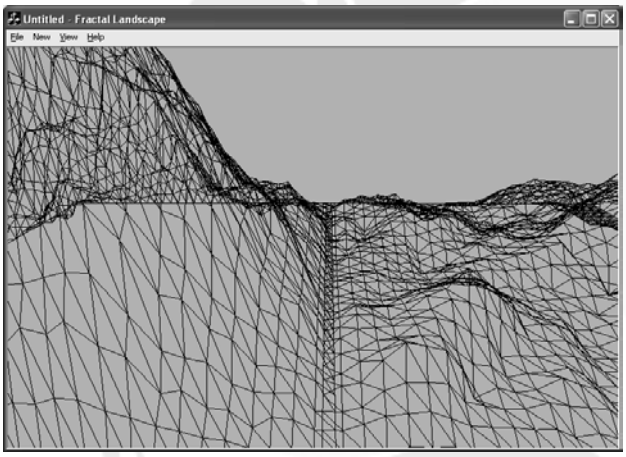


Figure 9. Mesh triangles of the terrain

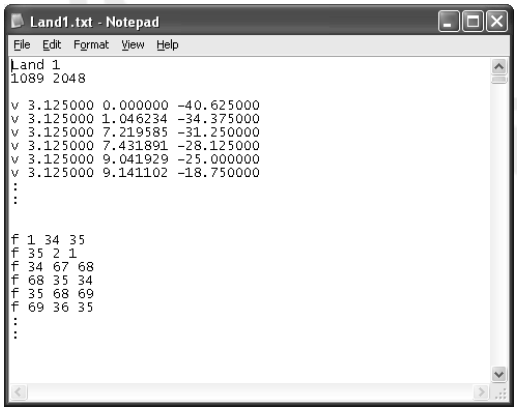


Figure 10. Data stored for a certain terrain

5. CONCLUSION

In general, this research can be said have already achieved its goal. Generating data to create terrain can be performed easily using this terrain generator. Users can create as many terrains as they need and like. The data of the terrains that are saved in a text file make it accessible and usable to any other programs.

From the visualization of the terrain, one can see that based on parameters provided, many variations of the terrains can be generated by the program. By applying the color and texture to the generated terrains, one can see the pleasing result of the terrains.

For further work, it could be better if the users can modify the generated terrains. Therefore the users can create the terrain that meets exactly to their preferences. Furthermore, the application can be made to allow users select their own preferred color or texture to be applied to the hills or valleys generated. All of this information then can be stored in a text file for further use in other applications.

6. REFERENCES

[1] Daniel, A. N.D. Random Midpoint Fractals. http://eldar.mathstat.uoguelph.ca/dashlock/ftax/RMP.html

[2] Donald, U. 2009. OpenGL Tutorial. http://www.swiftless.com/tutorials/opengl/opengltuts.html

[3] Eric. W. 1999. Gaussian Distribution. http://bbs.sachina.pku.edu.cn/Stat/Math_World/math/g/g084.htm

[4] Heinz, O., Hartmut, J., Dietmar, S. 2004. Chaos and Fractals: New Frontiers of Science. Springer.

[5] Taygeta. N.D. Generating Gaussian Random Number. http://www.taygeta.com/random/gaussian.html

[6] William, W., Paul, M., Forrest, M. 1997. The Fractal Landscape Realizer. http://research.esd.ornl.gov/realizer

[7] Wikipedia. 2009. Fractal Landscape. http://en.wikipedia.org/wiki/Fractal_landscape

[8] Wikipedia. 2010. Normal Distribution. http://en.wikipedia.org/wiki/Normal_distribution

[9] Wikipedia. 2010. Box-Muller Transform. http://en.wikipedia.org/wiki/Box–Muller_transform

[10] Wolfram. N.D. Box-Muller Transformation. http://mathworld.wolfram.com/Box-MullerTransformation.html

From Taiwan Puppet Show to Augmented Reality

Yang Wang
Takming University of Science and Technology
No. 56, Sec. 1, Huanshang Rd.
Neihu District, Taipei City, Taiwan
2-26585801, 2660. +886
yangwang@takming.edu.tw

Bo Ruei Huang
Takming University of Science and Technology
No. 56, Sec. 1, Huanshang Rd.
Neihu District, Taipei City, Taiwan
2-26585801, 2667. +886
andrew7595@yahoo.com.tw

Zih Huei Wang
Takming University of Science and Technology
No. 56, Sec. 1, Huanshang Rd.
Neihu District, Taipei City, Taiwan
2-26585801, 2667. +886
eva47301195@hotmail.com

ABSTRACT

In recent years Taiwan government has made the best effort in adding science technology and creative idea into Chinese traditional culture and arts to produce a new industrial and commercial product which is highly competitive on market, and can create considerable business opportunity over the world. Taiwan Puppet Show (Budaixi) [1]is a puppet controlled by hand. The stage for puppet performance is in the form of real stage but with reduced scale according to the size ratio of the real stage to human. The purpose of this research project is to virtualize the Taiwan Puppet Show. The research can be divided into two phases. In the first phase some rules were established by induction method, and in the second phase the computer interaction technique was introduced to develop application software which possesses the nature of the entertainment and value of traditional culture. This paper aimed at the second phase of the research project in which the detail of the design idea employed and the practice are introduced. The outcome of the research not only benefited the preservation of Chinese traditional culture, but also developed the "Finger Control Method" for "Augmented Reality" which can be employed on other complicated interactive operation.

Keywords

Computer graphic, Augmented Reality, simulation, Taiwan Puppet Show, hand control, human-computer interface.

1. INTRODUCTION

Taiwan Puppet Show was first known in Ming Dynasty of China, and was brought into Taiwan from Fujian at the later part of Ching Dynasty. It was then prosperous for almost 20 years. In few decades ago, the TV's Budai Show further joined audio and video, as well as laser technologies plus the movie shooting approach into the show with bizarre and complicate stories to replace the traditional show that teaches moral teachings. It had successfully attracted audience who are seeking for change all the time. The subordinated comic books, novels and figures followed. They are not only favored widely by the public in Taiwan, but also in USA and in part of European countries, they enjoy popularity to noticeable level.

Figure 1. Character in Taiwan puppet show

The purpose of the project is to create the roles which can be operated individually by multiple persons through computer network, and performed on the virtual stage. This research is supported by National Science Council of Taiwan, and the project is divided into two phases for execution. The first phase of the research is already completed. In the first phase the differentia of the relation between the traditional puppet operation method and the movement of fingers is determined through analysis. And, the complicated finger movement is simplified into the operation procedure of two degrees of freedom by taking the development of the tangible Interface as premise [2]. The second phase of the project is still in progress up to the present. In this phase the following problems are to be studied and solved:

- Based on the analysis result obtained from the first phase to determine the suitable finger movement, so that the features of the traditional puppet operation can be retained, and the virtual puppet can be easily operated by means of real time control.
- Before the development of the tangible interface, different operation interfaces shall be compared, and the most cost saving and operation effectiveness interface shall be determined.

- Through analysis to determine the suitable operation degree of freedom for the interface technique being developed, and obtain the trigger method for different puppet animations.

In the research the Augmented Reality technique was chosen, and the micro palm movement sensing method was developed. Hence after completion of this research, in addition to attain the goal of promoting traditional arts, the techniques of palm control system can be transplanted to different software through interface research to provide more diversified man-machine interaction model.

2. RELATED WORK

Before entering into the subject of the research, the related knowledge or research in two related fields must be understood. The first thing is to know the characteristics of Taiwan Puppet Show and the special features of its operation, and the second is to know the related research of Augmented Reality.

2.1 About Taiwan puppet show

In Section 1 of this paper the history of Taiwan Puppet Show has been introduced. In this section we will see the tricks of the operation method [1]. The puppet of Budaixi is controlled, and operated by palm. After fitting palm into the hollow part of puppet, the thumb and middle finger are employed to control the arms of puppet while the index finger is used to control the head of puppet, as shown in Fig 2. In traditional operation method the middle finger must be kept apart from the index finger an angle almost 90°, therefore the puppet may not be smoothly operated or occupational harm may be caused due to the difference in genetic characteristics of hand of different person. Owing to this and in order to enable the virtual drama to become popularized, the hand control of the puppet is arranged to have the index finger set apart from the middle finger a small angle in principle as shown in Fig. 3, on the other hand, since the head of the traditional puppet is made of wood carving, the expression can not be changed when there is change of emotion. Hence a series of limb movement is developed for representing the delicate change of emotion of human such as angry, bashful look, thinking and drunk etc.. In the following table 1 several examples are listed. These special features must be retained in this research so that the traditional culture can be imparted to the next generation through technology innovation.

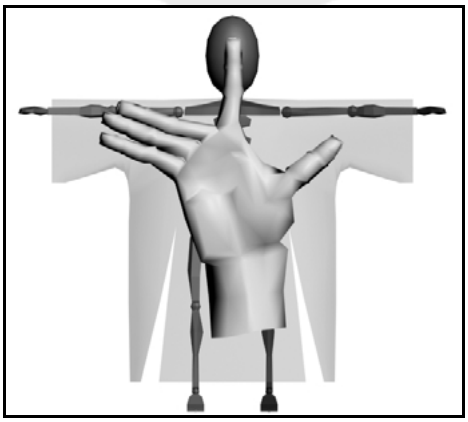


Figure 2. Standard puppet operation method [2]

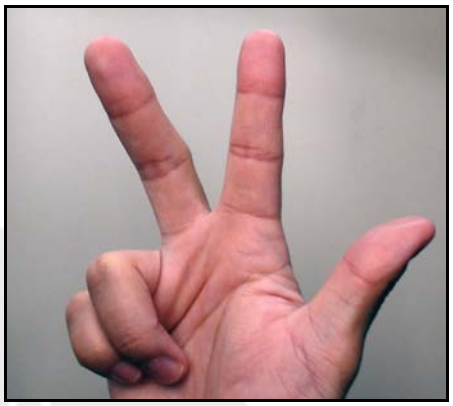


Figure 3. Small angle operation by middle finger and index finger

Table 1. Example of emotions and puppet movements [2]

Emotion	Motion 1	Motion 2	Motion 3
Angry	Shake head once heavily	Open arms	Bulge chest and belly
Furious	Put down hands straight	Shake head	Shake and lift body backward
Think	Head down a little bit	Knock gently with one hand	Body still
Modest	Put hands on belly	Bend waist a little bit	Nod repeatedly

2.2 Related research of Augmented Reality

Originally before starting the project, the operation method chosen for the research is the single degree of freedom bending detecting electronic element combined with wireless locating equipment which form the Tangible Interface. However, during the research it was found that the production cost of the operation glove was unable to be reduced, and the frequency of use of the bending detecting element is low, and it is not necessary to use the electronic part with such as high accuracy. Therefore the author of this paper has tried to seek other solution, and found that some of the related application methods of Augment Reality also meet the requirement of the research, but the production cost of the operation glove designed by employing the Augment Reality technology is low. Hence the author of this paper decided to introduce the AR technique into the research.

However since the Taiwan Puppet Show is operated by using palm and finger as controller which shall be developed by employing AR technique, the hand segmentation or 3D reconstruction for AR shall be chosen as the focal point of the preliminary research. Dr. Wu's research chose "easy to use" as the goal and "hand" as interface [3], and chose "skin detection" as focal point of research which has more degrees of freedom required in operation as based on the operation of puppet show, but will consume more system resource,

and will affect the display effectiveness of the 3D characters in the virtual stage. Hence it still needs to find the simpler and easier way. In the research of Kyungboo Jung a cubic with six different markers is employed to enable the virtual duplication of rigid objects from image sequence [4], the way employed by Kyungboo Jung is close to the design idea and needs of this research. Further, the author of this paper found that some combined detections are needed for designing the markers on fingers. The similar idea has been discussed in the research of Gun A, Lee [5]. And as for the design of marker detailed reference is included in the series study by H. Kato [6]. Besides, as for the execution of emotion movement high-level event system must be employed, Jean-Luc Lugrin has made a sample of practice [7], and an interface [8] has been developed and provided by R. Smith. It was also found during experiment that the position of the markers on the glove due to the operation of the glove is changed by different person. Hence as for the identification of marker the key point match method is adopted in this research [8].

3. STUDY OF THE OPERATION METHOD OF THE VIRTUAL CHARATER

3.1 Requirement of control interface

Here let us continue the discussion on the puppet operation method and characteristics of emotion movement etc., three principles for the design of operation interface can be firstly listed as follows:

- The detection of finger bending is achieved by single degree of freedom: during puppet operation, no matter which finger, thumb, index finger or middle finger, almost 99% of operations are in the direction of natural bending of finger knuckle (the only exception will be discussed in section 3.4).
- The hand only controls the single axis rotation and displacement: The turning round movement of puppet are achieved by having the middle finger as center shaft to turn left or right, while the walk motion of puppet is achieved mainly by the translation of the whole plane which is parallel to palm surface.
- The emotional expression is achieved by continuous movement: In the traditional puppet show the emotion is expressed by limb movement. In addition to the hand operated movement, the variation of face expression is triggered through the continuous signal on the virtual stage.

3.2 Discussion on operation degree of freedom

Since the hand operated puppet is designed according to the Limitation on the degree of freedom of fingers such as the whole piece of cloth is sewed on the shoulder joint so as to cover up the flaw that the finger can only bend in one direction. In addition, around each elbow of the puppet a large piece of hard cloth is sewed in order to simplify the operation of finger, and to avoid the strange feeling because the bending direction of puppet is different from the bending direction of the elbow of real man. Therefore, in the movement of puppet the arm is always kept in the position of stretch and extend foreword or hang downward close to body, so that in the arrangement of degree of freedom of finger, the bending angle of 90° is arranged as one graduation for the change of movement as shown in Figure 4. For example when the middle finger points to upward position the right arm of the puppet is in horizontal stretch position. When the middle finger rotate in counter clockwise direction for an angle of 90° the puppet is stretching forward its right arm in horizontal position, and when the middle finger continues to move in counter clockwise direction for 180°, the puppet put down its right arm. There is no other angular limb expression model required for arm movement..

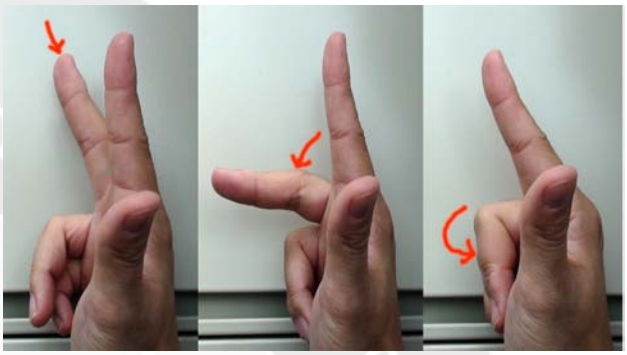


Figure 4. Movement of middle finger

Figure 5. Movements of puppet compared with the movements of finger shown in Figure 4

3.3 Design of marker

From Section 3.2 we see that a finger is equivalent to a cubic with multiple markers as shown in Fig. 6, three fingers are equivalent to 3 cubic, also add the palm into fingers, there are 4 control cubic. The control on the character by fingers is achieved in such a way that middle finger corresponds to right arm, index finger corresponds to head, and thumb corresponds to left arm. These three parts of control are not for displacement of puppet but are for animation, and the control achieved by palm is corresponding to the displacement of body, the movement of body will affect the movement of head and arms.

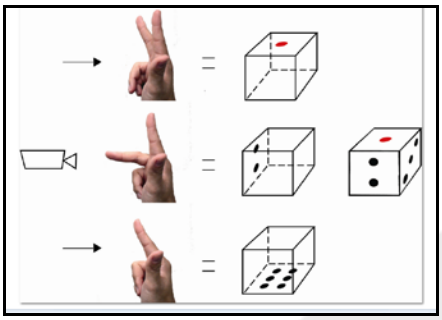


Figure 6. Markers on finger is like a cubic marker

The palm cubic is somewhat different from finger cubic. The palm cubic is mainly designed for the rotation positioning of puppet body. As based on the design point of view, since palm is in flat shape, we have compared the movement of the puppet with the rotation angle of palm viewed by camera, and observed that when palm is turning to left side, the marker on the center of palm will cause the puppet to turn, but when the palm turns to an angle greater than 50 degrees, it is apt to cause the character (virtual puppet) disappear which is possibly due to the curvature of palm. Owing to this we arranged a pair of markers on the edge of palm which can relocate the position and angle of the character when palm turns to this angular position as shown in Figure 7.

Figure 7. Markers on palm

When the palm is just facing the camera, the character is driven by Red Cross marker on the center of palm while the special markers are stuck on fingers. Again take middle finger as an example, shown in Figure 8 when three yellow markers are detected, it represents that the arms of the character are in open state (normal state). When the middle finger bends to an angle of 90 degrees, the markers will display an arrangement of Blue-Yellow-Yellow which represents that the right arm of the character will stretch and extend forward, and finally the arrangement is Green-Yellow-Yellow which represents that the right arm of the character will hand by the side, and close to body.

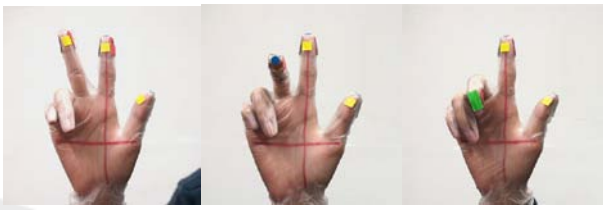


Figure 8. key point markers

When the palm turns with its side facing the camera, the marker on the side of finger will be employed to trigger the movement of character's limb. Let us continue the design idea stated above. The markers on one side of he fingers are all arranged in the same color for making sure the consistency of the character and the movement direction. In addition, the movement of fingers becomes even more directive if seen from the side of the fingers. Hence the geometric pattern was added to the color-marker to help identifying the specific finger and direction (0 degree – 90 degrees -180 degrees) as shown in Figure 9. The arrangement of thumb in the design is a special case, because thumb is even flatter in shape, and the included angle between the direction of thumb knuckle movement and palm surface is 45 degrees, therefore whether the character is in front position or side position, the same marker can be employed to achieve the control, however the identification in side position will lead to the directional difference of image.

Figure 9. Marker with geometric pattern

Besides, there is another special condition in the traditional script; the movements of scratching head are always included. This can be achieved simply by having two fingers placed closer to each other or open and close repeatedly). Hence the marker needs to have a special design which enables the effect of making the combined geometric pattern to from the third pattern. When the system detected the combined pattern shown in Figure 10, character will lift up its arm. If a repeated open-close of fingers is detected, the character will perform the movement of touching head as shown in the right side of the following figure.

Figure 10. Special designed marker

3.4 Operation theory for emotion and expression

In Table 1 some examples of emotion are listed. In Taiwan Puppet Show the emotion of character is represented by these limb movements. However in the virtual system we can use finger to operate the traditional continuous movement, and have the system to initiate expression through marker sequence, such as Figure 11. The example shown in Figure 12 is one of the state cycles for initiating emotion.

Figure 11. Face expressions (normal-happy-angry)

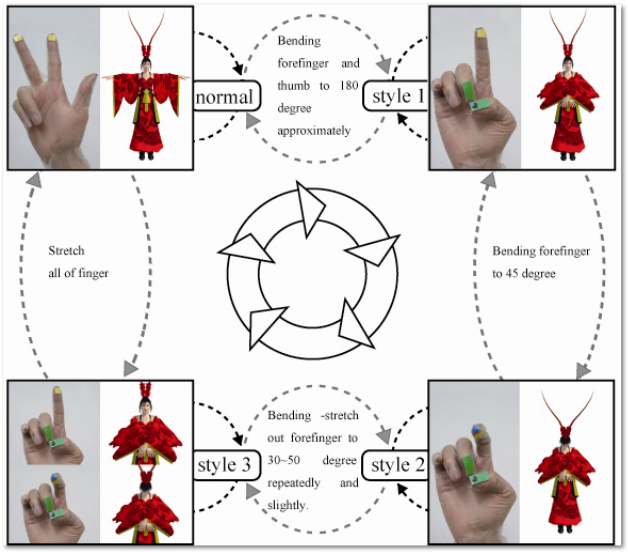


Figure 12. State of expression movement (modest)

4. IMPLEMENTATION

System of these researches is included in Intel Dual Core T4300 2.1GHz, 2GB DDRII RAM, Nvidia PCI-E Geforce G105M and 512MB. Some examples of the outcome of implementation are illustrated in Table 2.

Table 2. Examples of implementation
Top-down are normal, hang head, left hand stretch forward, left hand hang and close to body

5. CONCLUSION

The research is still in progress at present. The marker design mentioned in Section 3 and the prototype of emotion all were successfully completed. In the following months a lot of "emotion movement-expression" rule shall be completed. We will continue to improve the system in the future such as changing the morph of animation of character to the control of the Node in Web3D by Key point marker, so as to enable a more smooth performance of character, and hope the application of hand control system can be extended to the interface control of many more degrees of freedom.

6. ACKNOWLEDGMENTS

This research was supported by the National Science Council of Taiwan under the grant NSC98-2221-E-147-006.

7. REFERENCES

[1] Y.-B. Zhu, C.-J. Li, I. F. Shen, K.-L. Ma, A. Stompel, "A new form of traditional art: visualsimulation of Chinese shadow play," *Proceedings of the SIGGRAPH 2003 conference on Sketches & applications*, 2003.

[2] Yang Wang, "Analysis of character animation control in virtual Taiwan-Puppet Show", *IASTED conference: Advances in Computer Science and Engineering*, Thailand, 2009.

[3] Wu Yueming, He Hauwu, Ru Tong, Zheng Detao, "Hand segmentation for augmented reality system," *Second workshop on digital media and its application in museum and heritage*, 2007, 395-400.

[4] Kyungboo Jung, Seungdo Jeong, Byung-UK Choi, "Virtual duplication of rigid objects from image sequences," *International conference on computer science and software engineering*, 2008, 1166-1169.

[5] Gun A. Lee, Mark Billinghurst and Gerard Jounghyun Kim, "Occlusion based interaction methods for tangible augmented reality environments," *Proc. of ACM SIGGRAPH International Conference on Virtual-Reality Continuum and its Application in Industry*, 2004, 419-426.

[6] H. Kato and M. Billingghurst, "Marker tracking and HMD calibration for a video-based augmented reality conference system," In proceedings of the second IEEE and ACM international workshop on augmented reality, 1999, 85-94.

[7] Jean-Luc Lugrin, Remi Chaignon, Marc Cavazza, "A high-level event system for augmented reality," *IEEE/ACM International Symposium on Mixed and Augmented Reality*, 2007, 269-270.

[8] R. Smith, "Open dynamic engine," www.ode.org, 2005.

[9] Gerhard Schall, Helmut Grabner, Michael Grabner, etc., "3D tracking in unknown environments using on-line keypoint learning for mobile augmented reality," *In Proceedings Workshop on Visual Localization for Mobile Platforms*, 2008.

Generating Iriscode Using Gabor Filter

Darma Putra
Department of Electrical Engineering, Faculty of Engineering, Udayana University
duglaire@yahoo.com

Lie Jasa
Department of Electrical Engineering, Faculty of Engineering, Udayana University
liejasa@ee.unud.ac.id

ABSTRACT

Iris can be classified into the most reliable biometric characteristics because the variability of iris is very high. The important issue in iris biometrics recognition system is how to find the feature of iris. Gabor filter is used in this paper to produce the iris feature. This feature is called iris code. Hamming distance algorithm is used to find the similarity degree between tested and referenced image.

This system is tested using 300 iris images from 100 people which 2 samples as reference sample and 1 sample as experiment for the right eye, and 400 iris images from 100 users, 3 samples as reference and 1 sample as tested sample for the left eye. Iris codes that is produced by Gabor filter gives good performance to the iris recognition system with accuracy reached 94,758% (T = 0463, FNMR = 4500%, 0742% = FMR) for the right iris and 92,397% (T = 0470, FNMR = 6000%, FMR = 1603%) for the left iris. The experiment result also show that the performance of this system relatifely stable although the database size increased.

Keywords

Gabor filter, iris recognition, Hamming distance.

1. INTRODUCTION

Technological developments in the field of computer and information system have increased in line with developments in science and technology. Development in computer technology that many of the current study is one of biometric technology. Biometrika is one of the sciences of stating the characteristics of an individual and is unique to that individual. Technology for self-knowledge by using parts of the body or human behavior today has reached a remarkable development in the verification system to replace conventional. Various methods have been developed in the problems that arise, particularly the introduction of human identity. Basically every human being has something unique / special that only belongs to himself. This raises the idea to make it as a unique human identity. Facts must be supported by technology that can automatically identify / recognize a person with developments in the world using information and computer technology. Such recognition system is known as Biometrika Recognition System[8].

Utilization of the human body are unique to distinguish between each other, has proven to give results much more accurate in the identification process. One of the utilization of organs for identification is iris. Iris patterns remain unchanged after its formation period, ie when in the womb ages 3 to 8 months and is in a protected place that is not easily scratched or damaged[5]. Iris known to have a level of differentiation may be good enough to classify any individual by using iris patterns as a key pembedanya. Iris pattern recognition system is the most reliable system compared with other type biometrika[11]. This paper focuses on the implementation of the Gabor Filters for iris identification system.

2. COLLECTING DATA IRIS

This research was conducted by using samples taken from the CASIA (Chinese Academy of Sciences' Institute of Automation) eye Image databases. The eye images are segmented and normalized automatically by using method in [12].

3. FEATURES EXTRACTION

Iris feature is obtained using 2D Gabor filters. This filters can be produces using the equation below:

$$G(x,y,\theta,u.\sigma)=\frac{1}{2.\pi.\sigma^2}\exp\left\{-\frac{x^2+y^2}{2.\sigma^2}\right\}\exp\{2.\pi.i(u.x.\cos\theta+u.y.\sin\theta)\}$$ Where,

$i=\sqrt{-1}$

u is frequency of the sinusoidal wave.

θ is control of the orientation of Gabor function. σ is standard deviation of the envelope Gausian. x, y is coordinates of Gabor filters

Through the equation, will obtain a filter consist of real and imaginary parts filter. Gabor filters are implemented in the iris identification system has several parameters, both in terms of size filter (kernel), parameters θ (control of the orientation of Gabor Filters), parameters u (frequency of the sinusoidal wave), and the parameters σ (standard deviation of the Gaussian envelope) which adjusts filter size. Gabor filter parameters to be tested for the separation of the iris features can be seen in the following table.

Tabel 1. Gabor filter parameters

Size Filters	Angle (θ)	Sinusoidal frequency (u)	Standard deviation of Gaussian Envelope (σ)
3 x 3	-45^0	0.7332	0.7022
9 x 9	0^0	0.3666	1.4045
17 x 17	45^0	0.1833	2.8090
35 x 35	90^0	0.0916	5.6179

The normalized Gabor filters can be obtained by using the following equation.

$$\tilde{G}[x,y,\theta,u,\sigma] = G[x,y,\theta,u,\sigma] - \frac{\sum_{i=-n}^{n}\sum_{j=-n}^{n} G[i,j,\theta,u,\sigma]}{(2n+1)^2}$$

With $(2n+1)^2$ is the size of the Gabor filter. In fact, the imaginary part of Gabor filter automatically has zero DC because of odd size filters. It should be noted that the success of Gabor filter depend on parameters θ, σ, u selection for these filters.

Real and imaginary part of normalize Gabor filters are convoluted to the normalized iris images and produces real and imaginary characteristic features. Each pixel value of convolution result will be encoded into binary value (0 or 1) with rules bellow:

$br = 1$ if $Re\left[\tilde{G}[x,y,\theta,\sigma] * I\right] \geq 0$
$br = 0$ if $Re\left[\tilde{G}[x,y,\theta,\sigma] * I\right] < 0$
$bi = 1$ if $Im\left[\tilde{G}[x,y,\theta,\sigma] * I\right] \geq 0$
$bi = 0$ If $Im\left[\tilde{G}[x,y,\theta,\sigma] * I\right] < 0$

where *I* is normalize iris image and the operator * represents the convolution process.
Iris code is generated by the above rules to form a binary code which is indicated by bit 0 or 1 both for real and imaginary parts of Gabor filters (see Figure 1)

4. MATCHING

To obtain the similarity degree between test and reference images is used normalized Hamming distance as follow:

$$D_O = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}\left(P_R(i,j) \otimes Q_R(i,j)\right) + \left(P_I(i,j) \otimes Q_I(i,j)\right)}{2N^2}$$

With $P_R(Q_R)$ and $P_I(Q_I)$, respectively, expressed the real and imaginary parts of $P(Q)$. The results of the boolean operator $(\otimes)$ is equal to zero, if and only if the bits of $P_{R(I)}(i,j)$ equal to $Q_{R(I)}(i,j)$. The size of the matrix is given by NxN. The value of D_0 will have a range between 0 and 1.

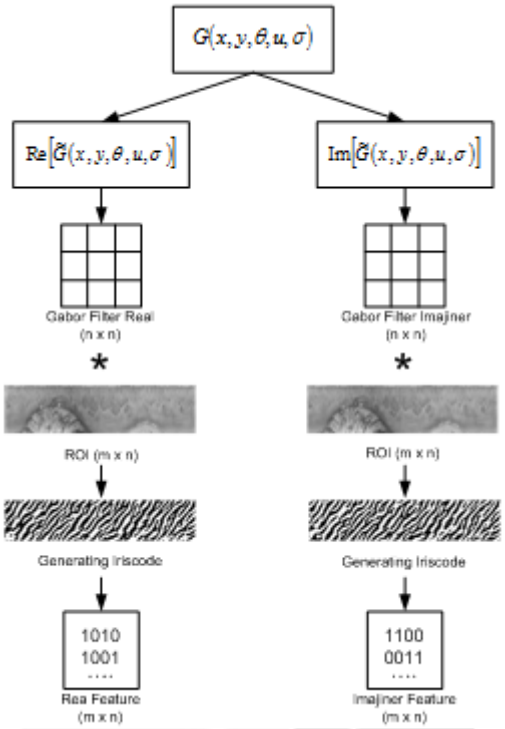
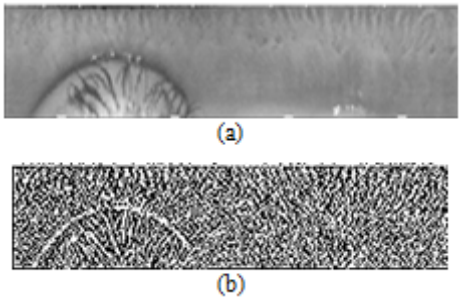


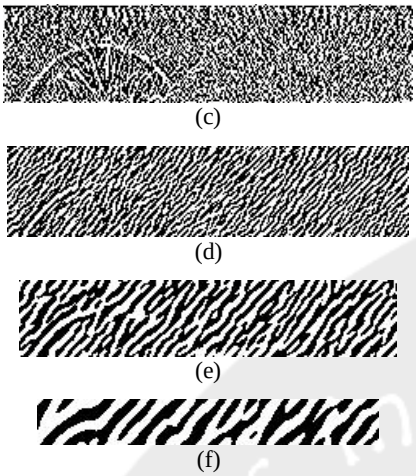
Figure 1. Block diagram of the formation process iriscode.

5. EXPERIMENT AND RESULT

The performance of identification system is measured by calculating the value FNMR, FMR or ERR. The testing is conducted in the form of simulation that compare the characteristics of each sample are listed in the database (reference sample characteristics with the characteristics of test samples).

Testing of this identification system using a database that contain 700 iris samples from 100 different people, which represented 1 of 3 slices of the sample right eye, and 4 slices of the sample left eye. Two of the three sample slices of the right side and three of the four sample slices of the left side are used as reference samples, while the rest is used as test samples, a total of 200 samples is a sample reference to the right eye, 300 samples for the left eye, and the total test samples is 100 samples for the right eye, and 100 samples for the left eye.
Separation of iris features with Gabor filter method, using the ROI slice image size 64 x 256 pixels. Here are presented the results of feature extraction of iris Figure 3 (a) with a different filter size (3x3, 5x5, 9x9, 17x17, and 35x35).





(c) (d) (e) (f)

Figure 2. Gabor features of a variety of filter sizes (a) is the ROI slices, (b) is a real feature of the filter size 3x3, (c) the characteristics of real filter size 5x5, (d) the characteristics of real filter size 9x9, (e) real characteristic filter size 17x17, (f) characteristics of real filter size 35x35.

The results of testing each Gabor Filters above can be summarized based on the level of the highest accuracy using the simulation as shown in Figure 3 and Figure 4. Simulation aims to find out the performance of identification systems based on the calculation of the value FNMR, FMR and the EER. User score legitimate (genuine scores) obtained by comparing the characteristics with the characteristics of the test iris reference iris of the same person, while the score is not a valid user (impostor scores) obtained by comparing the characteristics with the characteristics of the test iris reference iris that did not belong to the same person.

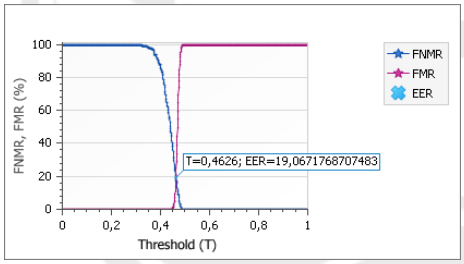


Figure 3. System Performance Characteristics curves using the filter size 3x3, angle 45 °, image slices of the ROI right eye 64x256 pixels and no translational.

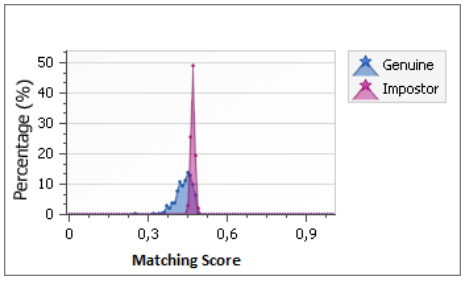


Figure 4. Probability Distribution curve Geniune User Score and Impostor using filter size 3x3, angle 45 °, image slices of the ROI right eye 64x256 pixels and no translational.

Table 2 show the accuracy level of Gabor Filters for the ROI slice of the right eye and Table 3 for the ROI slice of the left eye with a size of 64 x 256 pixels.

Table 2. ROI test results right iris 64x256 pixels, N=100, total image of the reference=200, total image of the test=100

Filter	Angle			
	-45^0	0^0	45^0	90^0
3 x 3 (%)	53.207	72.551	52.818	72.010
	79.556	94.495	78.904	77.576
	4 *pixel*	3 *pixel*	5 *pixel*	2 *pixel*
5 x 5 (%)	54.197	73.157	53.076	71.601
	80.697	**94.758**	81.298	80.364
	3 *pixel*	**3 *pixel***	4 *pixel*	3 *pixel*
9 x 9 (%)	40.182	45.843	37.702	43.020
	62.884	78.924	60.429	45.808
	4 *pixel*	4 *pixel*	4 *pixel*	4 *pixel*
17x17 (%)	58.369	68.399	56.247	49.056
	78.343	92.106	74.970	46.121
	3 *pixel*	4 *pixel*	4 *pixel*	3 *pixel*
35x35 (%)	72.879	83.838	69.172	61.278
	81.212	91.697	78.742	64.929
	4 *pixel*	3 *pixel*	5 *pixel*	8 *pixel*

Test results of Gabor Filters with slices of the ROI image of the right eye with a size of 64 x 256 pixels indicate the highest level of accuracy that is 94,758% obtained when using the filter size of 5 x 5 with angle 0 ° and translation of 3 pixels.

Table 3. ROI test results left iris 64x256 pixels, N=100, total=300 reference images, the total image of the test=100

Filter	Angel			
	-45^0	0^0	45^0	90^0
3 x 3 (%)	53.535	73.475	51.559	68.333
	74.070	**92.397**	70.095	75.377
	4 *pixel*	**8 *pixel***	4 *pixel*	4 *pixel*
5 x 5 (%)	55.199	73.754	56.047	69.159
	78.327	92.058	74.452	77.538
	3 *pixel*	3 *pixel*	6 *pixel*	8 *pixel*
9 x 9 (%)	37.444	47.829	38.148	36.471
	58.609	73.148	55.616	36.835
	2 *pixel*	5 *pixel*	4 *pixel*	9 *pixel*
17x17 (%)	58.260	69.411	53.374	43.287
	74.929	86.495	71.239	43.188
	3 *pixel*	3 *pixel*	4 *pixel*	9 *pixel*
35x35 (%)	68.643	83.084	65.057	59.619
	77.205	87.848	74.260	59.458
	8 *pixel*	2 *pixel*	8 *pixel*	9 *pixel*

Test results of Gabor Filters with slices of the ROI image of the left eye with a size 64 x 256 pixels indicate the highest level of accuracy that is 92,397% obtained when using the filter size 3 x 3 with angle 0 ° and translation of 8 pixels.

Table 4. Tapis test results for 5x5 ROI right iris 64x256 pixels on a variety of databases

Data base	(T) (%)	FNMR (%)	FMR (%)	Success Rate (%)
25	0.471	22.000	3.208	**74.792**
	0.461	4.500	0.563	**94.938**

	3pixel	3pixel	3pixel	**3pixel**
50	0.472	22.000	3.919	74.082
	0.462 3pixel	4.500 3pixel	0.602 3pixel	94.898 3pixel
75	0.473	22.667	4.883	72.450
	0.463 3pixel	4.667 3pixel	0.766 3pixel	94.567 3pixel
100	0.472	23.000	3.843	73.157
	0.463 3pixel	4.500 3pixel	0.742 3pixel	94.758 3pixel

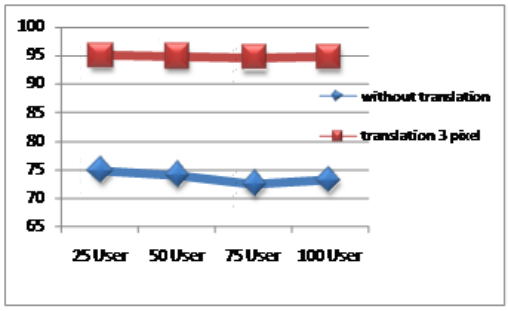


Figure 5. Accuracy Rate Graph Gabor Filters for the ROI of the right eye 64x256 pixels with different size databases.

The graph above show a linear line with the increasing number of users by using 3-pixel translation, in other words the level of accuracy as the system is stable enough to increase the number of users.

The graph in Figure 6 shows that the accuracy rate declined slightly and the system is stable enough with the addition of the number of users.

Table 5. Test results Filters 3 x 3 to the left of the ROI slice 64 x 256 pixels at various database sizes

Data base	(T) (%)	FNMR (%)	FMR (%)	Success Rate (%)
25	0.474	21.000	5.028	**73.972**
	0.469 8pixel	5.667 8pixel	1.250 8pixel	**93.084** 8pixel
50	0.476	21.667	4.973	73.360
	0.469 8pixel	6.000 8pixel	1.361 8pixel	92.640 8pixel
75	0.476	25.333	5.435	69.232
	0.470 8pixel	7.556 8pixel	1.658 8pixel	90.786 8pixel
100	0.475	22.333	4.192	73.475
	0.470 8pixel	6.000 8pixel	1.603 8pixel	92.397 8pixel

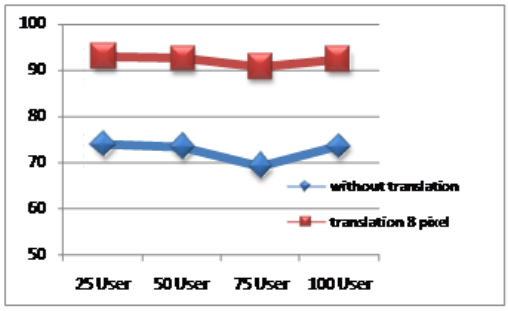


Figure 6. Accuracy Rate Graph Gabor Filters for the ROI of the left eye 64x256 pixels with different size databases

6. CONCLUSIONS AND FUTURE WORKS

Iris codes is produced by Gabor filter gives good performance to the iris recognition system with accuracy reached 94,758% (T = 0463, FNMR = 4500%, 0742% = FMR) for the right iris and 92,397% (T = 0470, FNMR = 6000%, FMR = 1603%) for the left iris. The experiment result also show that the performance of this system relatifely stable although the database size increased.This research just developed offline recognition system. The online iris recognition system from hardware device for acquisition image until sofware recognition system and applied the system in a public environment are research area for future.

7. REFERENCES

[1] Carr, D; Lipinski, B; Khabashesku, D. 2004. Iris Recognition: Gabor Filtering. (July.2008). DOI=hhtp:\cnx.org\contenm12493\lastest\.

[2] Chinese Academy of Sciences Institute of Automation. 2006. CASIA-IrisV3. (January, 2008).DOI=http:\\www.cbsr.ia.ac.cn/IrisDatabase.htm.

[3] Darma Putra, IKG. 2009. Sistem Biometrika Teori dan Aplikasi. Andi Offset : Jogjakarta

[4] Darma Putra, IKG. 2010. Pengolahan Citra Digital. Andi Offset : Jogjakarta

[5] Daugman, John. 2002. How Iris Recognition Works. Proceeding of 2002 International Conference on Image Processing. Vol.1.

[6] Fahmi. 2007. Perancangan Algoritma Pengolahan Citra Mata Menjadi Citra Polar Iris Sebagai Bentuk Antara Sistem Biometrik. Medan : Departeman Teknik Elektro Fakultas Teknik Universitas Sumatra Utara.

[7] Jahne, Bernd. 2002. Digital Image Processing. Jerman : Springer.

[8] Jain, Anil K., Bolle, Ruud., dan Pankanti, Sharath. 1999. Introduction to Biometrics. BIOMETRICS Personal Identification in Networked Society. London : Kluwer *Academic Publisher*.

[9] Jia, Chang. 2005. Automated Iris Recognition Technology & Iris Biometrc System.

[10] Kong, W K Adams. 2001. Using Texture Analysis on Biometric Technology for Personal Identification. Hong Kong : The Hong Kong Polytechnic University.

[11] Masek, Libor. 2003. Recognition of Human Iris Patterns for Biometrics Identification. The University of Western Australia.

[12] Nazer Jawas. 2009. Segmentasi dan Normalisasi Citra Iris. Bukit Jimbaran: Universitas Udayana.

Interpolation Technique to Improve Unsupervised Motion Vector Learning of Wyner-Ziv Video Coding

I M.O. Widyantara
Multimedia Communication Lab., Department of Electical Engineering, Institut Teknologi Sepuluh Nopember (ITS), Surabaya 60111, Indonesia
oka.widyantara@unud.ac.id

N.P. Sastra
Multimedia Communication Lab., Department of Electical Engineering, Institut Teknologi Sepuluh Nopember (ITS), Surabaya 60111, Indonesia
putra.sastra@unud.ac.id

D.M. Wiharta
Multimedia Communication Lab., Department of Electical Engineering, Institut Teknologi Sepuluh Nopember (ITS), Surabaya 60111, Indonesia
wiharta@unud.ac.id

Wirawan
Multimedia Communication Lab., Department of Electical Engineering, Institut Teknologi Sepuluh Nopember (ITS), Surabaya 60111, Indonesia
wirawan, gamantyo@ee.its.ac.id

G. Hendrantoro
Multimedia Communication Lab., Department of Electical Engineering, Institut Teknologi Sepuluh Nopember (ITS), Surabaya 60111, Indonesia
wirawan, gamantyo@ee.its.ac.id

ABSTRACT
One major problem in Wyner-Ziv video coding with unsupervised motion vector learning is on the quality of blockwise motion vector field that produced by motion estimator. Nearest neighbor interpolation of block-wise motion vector field onto pixelwise motion vector field in fact creates step-like transitions at block boundaries. The actual pixelwise motion vector field has smooth transitions across space. This paper proposes method of bilinear interpolation to improve pixelwise motion vector field which mitigates the boundary effects. The implementation of this method in WZ video codec with unsupervised motion vector learning has led to 3-4 % and 6-7% bit rate saving for GOP 2 and 8 respectively, when compare to nearest-neighbor interpolation method. Rate-distortion performance improved to 0.46 dB and 0.27 db for GOP 2 and 8 respectively.

Keywords
WZ video coding, Nearest-neighbor interpolation, Bilinear interpolation, EM algorithm.

1. INTRODUCTION
Wyner-Ziv video coding offers an alternative to predictive video coding methods for applications requiring low-complexity encoding, such as wireless low-power video surveillance and video sensor networks. Wyner-ziv video coding limits the encoding complexity to a level at which video frames can be encoded separately, not jointly. Separately encoded frames can nevertheless be decoded jointly, since there is no complexity constraint at the decoder. Specifically, each Wyner-Ziv encoded frame can be decoded with reference to side information derived from one or more already reconstructed frames. The theoretical basis of Wyner-Ziv coding for video was proposed in 1970s by Slepian-Wolf [1] and Wyner-Ziv [2]. These theorems mainly state that separate encoding and joint decoding of two correlated sources, *X* and Y, can be as efficient as joint encoding and decoding as in conventional video coding. The Slepian-Wolf theorem refers to lossless compression while the Wyner-Ziv theorem refers to lossy compression of X with side information Y available at the decoder.

Based on those theorems [1,2], Wyner-Ziv video coding separates encoding of frames precludes motion estimation at the encoder. The decoder, instead, estimates the motion in order to construct appropriately motion-compensated side information. Therefore, the quality of the side information plays a central role in the Wyner-Ziv (WZ) codec's overall rate-distortion (RD) performance. Without a powerful side information creation mechanism, no decent RD performance can be achieved. This fact motivated the development of many improvements in the simple and inefficient early side information methods. A novel and effective system for Wyner-Ziv (WZ) coding of video is presented in [3], with proposed a mechanism for unsupervised motion learning of forward motion vectors during the decoding of a frame using a expectation maximization algorithm. As shown in Figure. 1, the background WZ coding architecture in this work uses Low-Density Parity-Check (LDPC) codes for Slepian-Wolf coding and works at the symbol level using joint bitplane decoding. This option is justified with the need to obtain an accurate motion field distribution. This motion learning framework with the "soft"(statistical) value at the LDPC decoder output to update the motion field, which is used by probability model to update the "soft" side information that fed to the LDPC decoder.

The performance of method [3] showed that the RD performances are lower than Wyner-Ziv video codec with motion oracle. This showed that the motion estimator process still put on differences between WZ frame (*X*) and the shifted version of *Y* (side information frame). Error remaining after forward motion vector field-compensation is modeled by additive noise. Thus, the overall relationship between X and Y is

$$X(x,y) = Y\big(x - M(x,y), y - M(x,y)\big) + N(x,y) \qquad (1)$$

where N is modeled as Laplacian-distributed noise independent of X, Y, and M.

Estimation from motion vector field M acquired by using block matching algorithm, that is computed based on block by block size k-by-k as such that noise N can be assumed to be stationary. Simplification in this motion estimation can be though as that all the pixels in a given block share the same $P(M)$. Block-based motion estimation produces blockwise constant motion vectors.

In this paper, we present an improvement method in the accuracy of estimation of motion vector field M in motion learning base Wyner-Ziv video codec [3] by interpolation of a block by block motion vector field onto pixel by pixel motion vector field such that allowing the noise N to be nonstationary across image. Based on previous work [3,4], we implement bilinear interpolation technique for pixel by pixel motion vector field, which increases the resolution of the motion vector field and achieves smoother motion field transitions.

2. THE WYNER-ZIV VIDEO CODEC

The tool proposed in this paper are implemented and evaluated in the context of the state-of-the art transform domain unsupervised forward motion vector learning Wyner-Ziv video codec [3], illustrated in Figure 1.

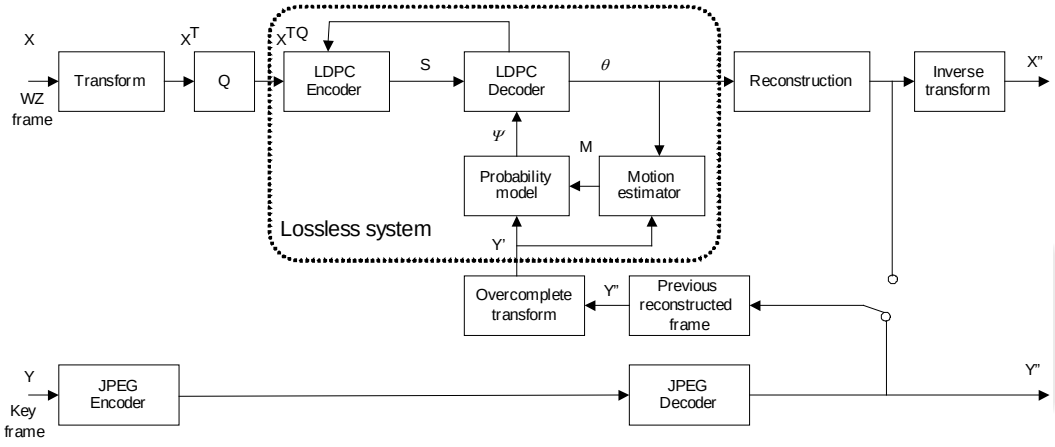


Figure 1. WZ video codec architecture with unsupervised forward motion vector learning

One frame Y (key frame) is transmitted by conventional coding, such as JPEG. Another frame X (Wyner-Ziv frame) must be encoded independently from Y but be decoded using Y' as side information. To exploit spatial correlation, the Wyner-Ziv encoder transforms X into *8-by-8* blockwise discrete cosine transform (DCT) coefficients X^T and quantizes X^T by a JPEG-recommended quantization table. Quantized coefficients X^{TQ} are losslessly communicated using a rate-adaptive low-density parity-check (LDPC) code [5]. The rate adaptive LDPC code enables small portions of the syndrome S to be incrementally sent, assuming a feedback channel is available.

The Slepian-Wolf decoder within the larger Wyner-Ziv decoder in Figure 1 is the loop formed by the LDPC decoder, the motion estimator and the probability model side information generator. This loop is an instance of the EM algorithm, as detailed thoroughly in figure 2. Using the received portions of S and the reference image Y', the Slepian-Wolf decoder iteratively estimates the quantized transform coefficients X^{TQ}. When the motion field M between X and Y' can be accurately estimated, the Slepian-Wolf decoder provides significant bit rate saving over conventional lossless transmission of X^{TQ}. Thus, reliable motion field estimation is a critical step in efficient coding.

As part of the decoder's EM algorithm, the E-step updates the estimation distribution on motion field M with block-by-block motion vector $M_{u,v}$. For a specified blocksize k, every k-by-k block of $\theta^{(t-1)}$ is compared to collocated block of Y as well as those ini a fixed motion search range around it. For a block $\theta_{u,v}{}^{(t-1)}$ with top left pixel located at (u,v), the distribution on the shift $M_{u,v}$ is update as below and normalized :

$$P_{app}^{(t)}\{M_{u,v}\} := P_{app}^{(t-1)}\{M\}P\left\{Y'_{(u,v)+M_{u,v}} \mid M_{u,v}; \theta_{u,v}^{(t-1)}\right\} \qquad (2)$$

where $Y'_{(u,v)+M_{u,v}}$ is the $k-by-k$ block of Y' with top left pixel at $((u,v)+Mu,v)$. Note that $P\left\{Y'_{(u,v)+M_{u,v}} \mid M_{u,v}; \theta_{u,v}^{(t-1)}\right\}$ is the probability of observing $Y'_{(u,v)+Mu,v}$ given that it was generated through vector $M_{u,v}$ from $X_{u,v}$ as parameterized by $\theta_{u,v}^{(t-1)}$. This procedure, shown in the left of Figure 2, occurs in the motion estimator.

The M-step updates the soft estimate θ by maximizing the likelihood of Y' and syndrome S.

$$\begin{aligned}\theta^{(t)} &:= \arg\max_{\Theta} P\left\{\hat{Y}, S; \theta\right\} \\ &= \arg\max \sum_{m} P_{app}^{(t)}\{M = m\}P\left\{\hat{Y}, S \mid M = m; \Theta\right\}\end{aligned} \qquad (3)$$

where the summation is over all configurations m of the motion field. True maximization is approximate by generating soft side information $\psi(t)$, followed by an iteration of joint bit-plane LDPC decoding to yield $\theta^{(t)}$.

The block-wise a posteriori distribution of the motion $P_{app}^{(t)}\{M_{u,v}\}$ weights the estimates from each of the blocks $Y'_{(u,v)+Mu,v}$, which are then summed into soft side information $\psi_{u,v}^{(t)}$, as shown in the

probability model step in the right of Figure 2. The probability that the blended side information has value w at pixel *(i, j)* is

$$\psi^{(t)}(i,j,w) = \sum_m P_{app}^{(t)}\{M=m\}P\{X(i,j)=w \mid M=m, Y'\}$$
$$= \sum_m P_{app}^{(t)}\{M=m\}PZ\left(w - Y_m^{'}(i,j)\right) \quad (4)$$

where $P_Z(z)$ is the probability mass function of the independent additive noise *Z*, and Y'_m is the previous reconstructed frame compensated through motion configuration *m*.

After the coefficients X^{TQ} are recovered by the Slepian-Wolf decoder, the Wyner-Ziv decoder proceeds to reconstruct the actual image. The existing one-disparity system uses nearest-neighbor reconstruction [6].

3. MOTION FIELD INTERPOLATION METHODS

The equation in (2), makes a simplification in motion vector field estimation by calculating the distribution $P(M)$ only on a block-by-block basis. If the frame is subdivided into square blocks of size *k*-by-*k*, then all the pixels in a given block share the same $P(M)$. The Interpolation technique is used for smooth transitions across space in the blockwise motion vector field.

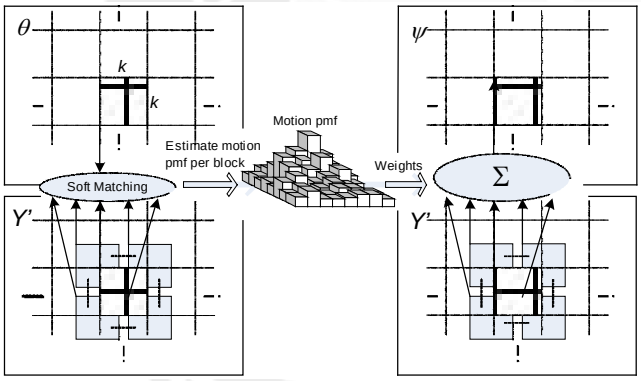


Figure 2. E-step motion estimation (left) and probability model (right) for Expectation Maximization algorithm.

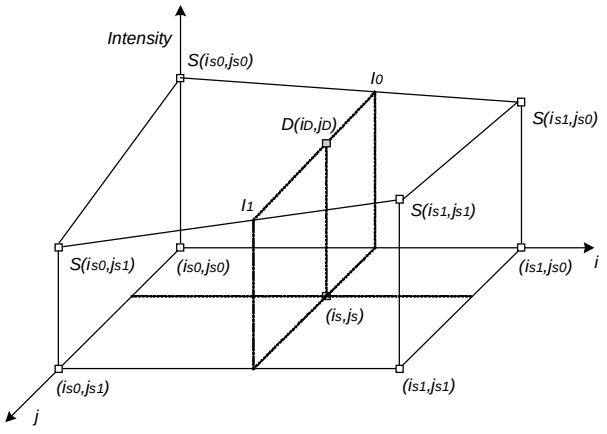


Figure 3. Bilinear interpolation

3.1 Nearest-Neighbor Interpolation

Using a single distribution $P(M)$ for all the pixels in a *k-by-k* block is equivalent to nearest neighbor interpolation of a small motion vector field onto a larger motion vector field. Specifically, if $M_{(u,v)}$ denotes the *block-by-block* motion vector field and $M_{(i,j)}$ denotes the *pixel-by-pixel* motion vector field, then

$$P(M(i,j)) = P\left(M\left(\lfloor i/k \rfloor, \lfloor j/k \rfloor\right)\right) \quad (5)$$

where, the interpolation is performed for each value *M(i,j)* can take. But, as explained in [4], nearest neighbor interpolation of a block-by-block motion vector field onto a *pixel-by-pixel* motion vector field can yield unnatural step-like transitions.

3.2 Nearest-Neighbor Interpolation

The linear interpolation algorithm uses source block intensities at the four pixels (i_{S0}, j_{S0}), (i_{S1}, j_{S0}), (i_{S0}, j_{S1}), (i_{S1}, j_{S1}) that are closest to (i_S, j_S) in the source block. As shown in Figure 3, at first, the intensity values are interpolated along the i-axis to produce two intermediate results I_0 and I_1.

$$I_0 = S(i_S, j_{S0}) = S(i_{S0}, j_{S0})*(i_{S1} - j_S) + S(i_{S1}, j_{S0})*(i_S - j_{S0})$$
$$I_1 = S(i_S, j_{S1}) = S(i_{S0}, j_{S1})*(i_{S1} - j_S) + S(i_{S1}, j_{S1})*(i_S - j_{S0}). \quad (6)$$

Then, the sought-for intensity $D(x_D, y_D)$ is computed by interpolating the intermediate values I_0 and I_1 along the *j*-axis:

$$D(i_D, j_D) = I_0*(j_{S1} - j_S) + I_1*(j_S - j_{S0}) \quad (7)$$

Where, (i_D, j_D) is pixel coordinates in the destination block, (i_S, j_S) is the computed coordinates of a point in the source block that is mapped exactly to (i_D, j_D), $S(i,j)$ is pixel value (intensity) in the source block and $D(i,j)$ is pixel value (intensity) in the destination block.

4. EXPERIMENTAL RESULTS

The results presented in this section are based on simulation ran of transform-domain unsupervised motion vector learning Wyner-Ziv video codec [3] with the video sequence QCIF-size Foreman at 15 fps and only the first 40 frames incorporated in this simulation. These sequences are chosen because they have already used in the previous work.

In simulation, domain transform performed by blocksize $k = 8$ for DCT. Quantization matrix adopts a scaled version of the one in Annex K of the JPEG standard with scaling factors Q = 0.5, 1, 2, and 4 [7]. The LDPC encoder using a regular degree 3 LDPC accumulate code with length 50688 bit as a platform for joint bitplane LDPC decoding with bit depth d = 8.

At this setting, exactly 6336 transform coefficients can be Wyner-Ziv coded at a time. For this reason, each QCIF-sized Wyner-Ziv frame divide into quadrants and code each quadrant separately using corresponding quadrant of previous reconstructed frame as decoder reference.

Interpolation of a *block-by-block* motion vector field onto a pixel-by-pixel motion vector field uses the shift range ± 10 pixels. This associates with the number of motion search range ± 10 pixels horizontally and vertically that used in estimation of motion vector field at the motion estimator and the probability model. The EM algorithm at the decoder that is initialized with a good value for the

variance of the Laplacian noise and the value of distribution for motion vector $M_{u,v}$, we adopted from [3], i.e :

$$P_{app}^{(0)}\{M_{u,v}\} := \begin{cases} \left(\frac{3}{4}\right)^2, & \text{, if } M_{u,v} = (0,0) \\ \frac{3}{4}\cdot\frac{1}{80}, & \text{if } M_{u,v} = (0,*),(*,0) \\ \left(\frac{1}{80}\right)^2, & \text{otherwise} \end{cases} \quad (8)$$

After 50 iterations of EM, if the reconstructed *X'* still does not satisfy the syndrome condition, the decoder request additional syndrome bit from decoder via feedback channel. The analysis on implementation of interpolation method in transform domain unsupervised motion vector learning Wyner-Ziv video codec [3] done by comparing nearest neighbor interpolation method and bilinear interpolation at GOP 2 and 8, respectively. Comparison with zero motion coding scheme and motion oracle also held with following previous work [3]

4.1 Bit rate saving for lossy coding

The main target of the implementation bilinear interpolation in the motion learning based WZ video codec are the production of more accurate soft side information for the LDPC decoder. Therefore, after decoding the sequences with the proposed interpolation method approach, it is expectable to observe bit rate savings from average value of decoded frame regarding the motion learning based WZ video codec

Tabel 1. Comparison of bit rate saving of motion learning based WZ video codec using interpolation methods

Interpolation Methods	Rate of motion learning base WZ video codec with GOP 2 (kbps)				Rate of motion learning base WZ video codec with GOP 8 (kbps)			
	Q = 0,5	Q = 1	Q = 2	Q = 4	Q = 0,5	Q = 1	Q = 2	Q = 4
Nearest-Neighbor	546	384	Q = 0	Q = 1	517	363	234	151
Bilinear	524	370	247	169	481	338	217	142
Saving rate	4%	4%	4%	3%	7%	7%	7%	6%

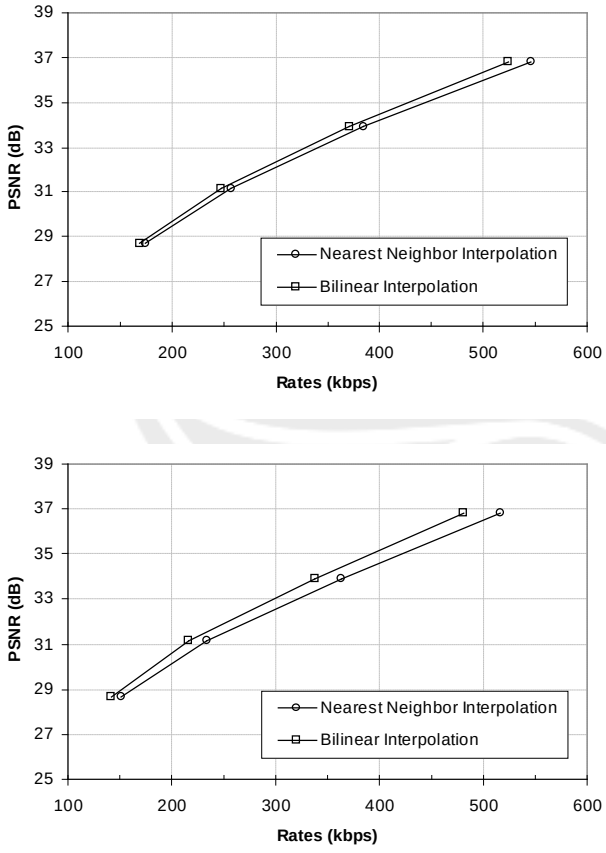


Figure 5. RD performances for motion learning based WZ video codec with different interpolation method for GOP 2 (above) and 8 (below)

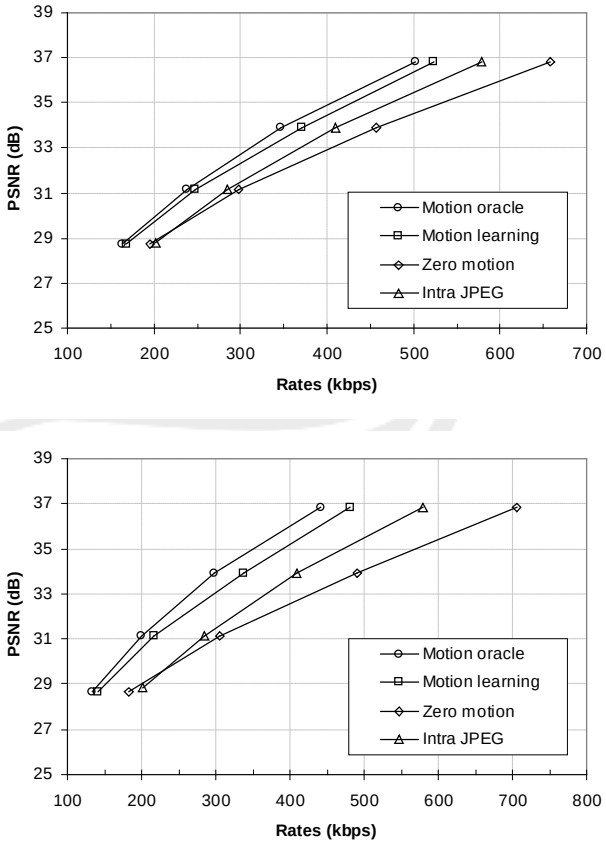


Figure 6. The comparison of RD performances of WZ video coding scheme with motion learning, motion oracle and zero motion, when bilinear interpolation conducted. GOP = 2 (above) and GOP = 8 (below)

Tabel 1. shows average bitrate yielded by motion learning based WZ video codec to code WZ frame using nearest-neighbor interpolation and bilinear interpolation. The bitrate is taken using JPEG quantization matrix with scaling factor Q = 0,5, 1, 2 and 4.

The bilinear interpolation method reduced bit rate 3-4% for GOP2 and 6-7% for GOP 8, compared with nearest-neighbor interpolation method. These results show that bilinear interpolation capable in increasing motion field resolution and mitigates the boundary effects and provides smooth transitions between blocks.

4.2 Rate – Distortion Performance

Figure. 4 show performance comparison of rate-distortion motion learning based WZ video codec when bilinear interpolation and nearest-neighbor interpolation conducted. Overall, at specified rate, bilinear interpolation method able to deliver better coding quality. For GOP 2, bilinear interpolation method yield average PSNR value of 0.46 dB higher than nearest-neighbor method, and 0,27 dB of GOP 8.

Figure. 5 show the comparison of R-D performances of WZ video codec scheme with motion learning, motion oracle and zero motion as conducted at previous work [3]. In this comparison, all WZ video codec scheme uses bilinear interpolation to construct interpolation from *block by block* onto *pixel by pixel*. As shown in Figure 5, GOP 2, WZ video codec with motion learning, its performance is lower than WZ video codec with motion oracle, 0.4 dB to be exact, but still 1.4 dB above WZ video codec with zero motion reference codec. For GOP 8, WZ video codec with motion learning, 0.6 dB lower than WZ video codec with motion oracle and 2 dB above WZ video codec with zero motion one.

5. CONCLUSIONS

This paper has proposed an approach to improve interpolation technique from block-by block motion vector field onto pixel-by-pixel motion vector field in Wyner-Ziv video codec with unsupervised forward motion vector. Method of bilinear interpolation improves motion vector field quality that yielded by motion estimator.

As compare to nearest-neighbor interpolation method, the bilinear interpolation is able to deliver bitrate saving 3-4% for GOP 2 and 6-7% for GOP 8. Rate distortion performances of motion learning base WZ video codec also show that bilinear interpolation result in 0.46 dB and 0.27 dB higher than nearest-neighbor interpolation for GOP 2 and 8 respectively. When bilinear interpolation used in every WZ video codec scheme, RD performances of motion learning base WZ video codec score below motion oracle base WZ video codec about 0.4 dB and 0.6 dB for GOP 2 and 8 respectively.

Motion field estimation in E-step of EM algorithm uses fixed size 8 x 8. Further works, in search of having true motion within blocks, varying block size is an alternative way to mitigate block artifacts. Variation in the block size *k* explores a tradeoff between motion field resolution and noise rejection.

6. ACKNOWLEDGMENTS

We would like to thank Mr. David Varodayan and David Chen of Information System Lab., Department of Electrical Engineering at Stanford University for sharing the code of Wyner-Ziv video codec.

7. REFERENCES

[1] D. Slepian and J.K. Wolf, “Noiseless coding of correlated information sources,” IEEE Trans. Inform. Theory, vol. IT-19, pp. 471-480, July 1973

[2] A. D. Wyner and J. Ziv, “The Rate-Distortion Function for Source Coding with Side Information at the Decoder,”IEEE Transactions on Information Theory, vol. 22, no. 1, pp.1–10, January 1976.

[3] D. Varodayan, D. Chen, M. Flierl and B. Girod, “Wyner-Ziv coding of video with unsupervised motion vector learning,” EURASIP Signal Processing: Image Communication Journal, Special Issue on Distributed Video Coding, vol. 23, no. 5, pp. 369-378, June 2008.

[4] D. Chen, D. Varodayan, M. Flierl, and B. Girod, "Distributed stereo image coding with improved disparity and noise estimation", *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Las Vegas, Nevada, March 2008

[5] D. Varodayan, A. Aaron and B. Girod, Rate-adaptive distributed source coding using low-density parity-check codes, *Proc. Asilomar Conference on Signals, Systems, and Computers, 2005*, Pacific Grove, California, November 2005

[6] B. Girod, A. Aaron, S. Rane, and D. Rebollo-Monedero, “Distributed video coding,” *Proc. IEEE*, vol. 93, no. 1, pp. 71–83, Jan. 2005.

[7] ITU-T, I. JTC1, Digital compression and coding of continuous-tone still images, ISO/IEC 10918-1 — ITU-T Recommendation T.81 (JPEG)

Iris Segmentation and Normalization

Darma Putra
Department of Electrical Engineering, Faculty of Engineering, Udayana University
duglaire@yahoo.com

Piarsa
Department of Electrical Engineering, Faculty of Engineering, Udayana University
piarsa@ee.unud.ac.id

Nazer Jawas
Department of Electrical Engineering, Faculty of Engineering, Udayana University
nazerjawas@yahoo.com

ABSTRACT

Iris segmentation is important issue in iris recognition system. The aim of iris segmentation is to obtain the ROI (region of interest) iris from the eye image. This paper compares two methods of iris segmentation, namely: Daugman and Hough method. Some image preprocessing techniques are applied to improve the performance of these methods. The next step after iris segmentation is normalization process. The normalization transforms the iris image in Cartesian coordinates to the polar coordinates. The images are obtained from the normalization will be used in iris features extraction stage.

The performances of those methods are tested using eye images from database CASIA 1, 2 and 3 databases. The experiment results show that the Hough method is better than Daugman. The best accuracy of Daugman method achieves 92.57%, 26.75%, 56.03% for the CASIA 1, CASIA 2 and CASIA 3 respectively, and the best accuracy of Hough method for the same databases is 96.57%, 54.5%, 94.83% respectively.

Keyword

Daugman, Hough, iris segmentation, iris normalization.

1. INTRODUCTION

Iris can be classified into the most reliable biometric characteristics because the variability of iris is very high [9]. In addition, the iris patterns also remain unchanged after the period in which the establishment is in the womb ages 3 to 8 months. Iris is also in a protected place that is not easily scratched or damaged [6].

The important issues in iris biometric recognition are iris acquisition, segmentation and normalization, feature extraction and matching. This paper is only focus on iris segmentation and normalization.

The aim of iris segmentation is to extract the ROI of iris from the eye image automatically. The result of segmentation process will be determines the performance of overall iris recognition system because the unique patterns (features) of iris will be obtained from the ROI iris image.

This paper compares two methods for iris segmentation, namely: Daugman and Hough methods. To improve the performance of those methods we perform some image preprocessing techniques.

After image segmentation, the normalization processes is applied to segmented iris. The normalization transforms the iris image in Cartesian coordinate to the polar coordinate. The normalized iris image will be used in next step of iris recognition system, namely iris features extraction.

2. COLLECTING IRIS IMAGE

This paper use source of data from the eye image databases of Chinese Academy of Sciences Institute of Automation (CASIA). CASIA databases image was taken using a special infrared camera designed for biometrics research purposes. CASIA databases have 3 different versions with different characteristics in each version. CASIA 1 is taken from a short distance without the camera lights. CASIA 2 is taken from a little distance away from the CASIA 1 and is taken without using the camera lights, while CASIA 3 was taken from a short distance using the camera lights [2].

The following are examples of the eye image from CASIA database.

(a) (b) (c)

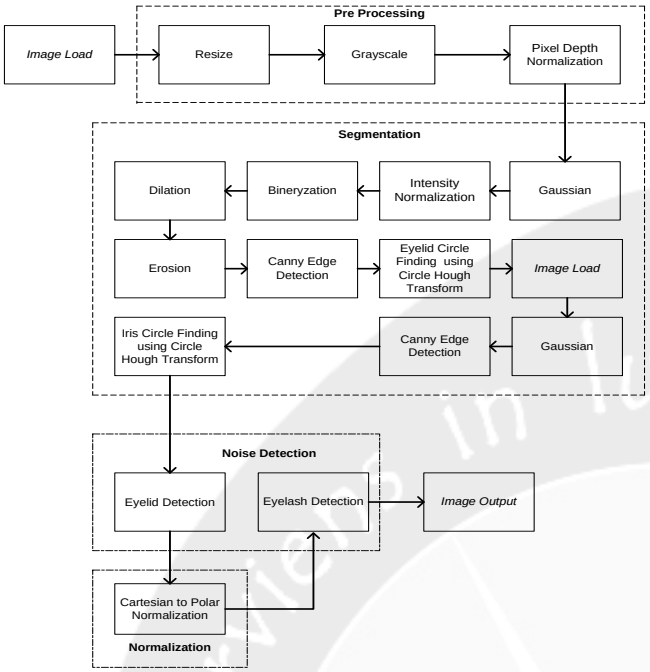
Figure 1. (a) (b) (c) example of eye image from CASIA 1, 2 and 3 database respectively.

3. SYSTEM OVERVIEW

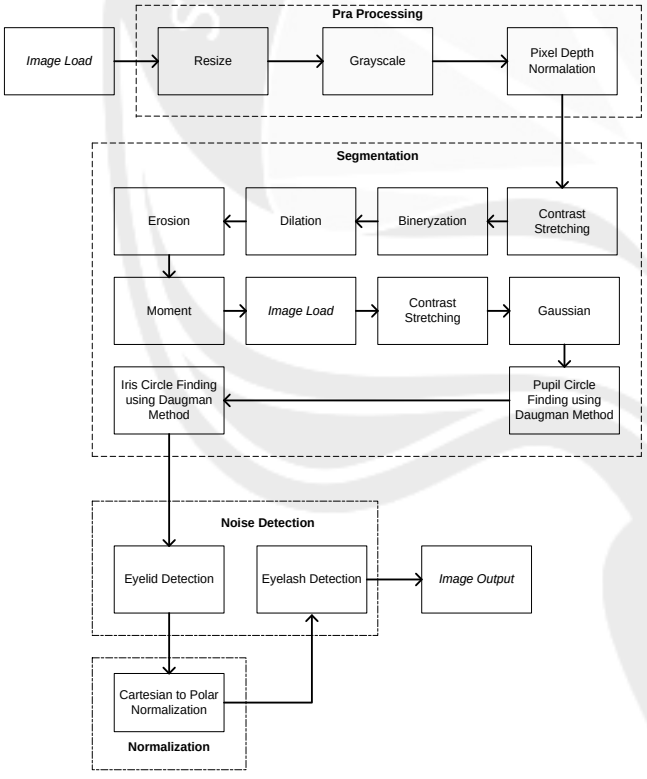
Figure 2 shown segmentation system overview of Daugman and Hough methods.
In generally, the segmentation systems consist of 4 subsystems, namely: image preprocessing, segmentation, noise detection and normalization. Both Daugman and Hough methods, the process of preprocessing, noise detection and normalization are same.

4. IMAGE PREPROCESSING

Image preprocessing consist of 3 process: image resizing to obtain the uniform image size (256 pixel in width and the high of image is adjusted to the original image height), gray scaling to obtain gray scale image and gray scale pixel depth changing to obtain the same depth pixel resolution.



(a)



(b)

Figure 2. Segmentation system overview, (a) Hough, (b) Daugman method

5. HOUGH SEGMENTATION

Steps of Hough methods are: Gaussian filtering, intensity normalization, image thresolding, dilation, erosion, edge detection, pupil circle seeking using Hough, image loading, Gaussian filtering, Canny edge detection, and iris circle seeking using Hough respectively.

5.1 Gaussian Filter

Gaussian process serves to blur (smooth) the image and reduce the noise. Gaussian filter (kernel) is obtained by using Gaussian formula bellow:

$$G(x,y) = \frac{1}{2\pi\alpha^2}\exp\left(-\left(x^2+y^2\right)/2\alpha^2\right) \quad (1)$$

5.2 Intensity Normalization

The intensity normalization is needed to reduce the possible imperfections in the palmprint image due to non-uniform illumination. The normalization process can be done as follow:.

$$I'(x,y) = \begin{cases} \phi_d + \lambda & \text{if } I(x,y) > \phi \\ \phi_d - \lambda & \text{otherwise} \end{cases} \quad (2)$$

$$\lambda = \sqrt{\frac{\rho_d\{I(x,y)-\phi\}^2}{\rho}} \quad (3)$$

where I and I' represents original gray scale iris image and the normalized image respectively, ϕ and ρ represents mean and variance of the original image respectively, while ϕ_d and ρ_d represents the desired values for mean and variance respectively. This research use ϕ_d = 100 and ρ_d = 100 for all experiments.

5.3 Image thresolding

The gray scale iris image is thresholded to obtain the binary iris image..

5.4 Dilation

Dilation morphology is a process to expand the foreground (object) area. Foreground pixels are noted as black pixels and background pixels are white pixels. The steps of dilation process as follows:

1. For each image pixel value do the following:
2. Check whether the neighbor pixels are black pixels (value (0)). If yes, then set the current pixel value with value 0. This process uses 8-pixel neighborhood.

5.5 Erosion

Erosion is the opposite of dilation. If the dilation process produces image with wider object area, the erosion produces image with smaller object area than the original image version. Steps of erosion process as follows:

1. For each image pixel value do the following:

2. Check whether the surrounding pixels (pixels neighbor) have a white pixel value (255). If yes, then set the pixel value with a white value (255). This process uses 8-pixel neighborhood.

5.6 Edge detection

This paper used Canny edge detection to find the edge iris image. Canny edge detection has very good ability to perform edge detection. This method can detect edges image in various directions, not just limited to horizontal or vertical edge. The purposes of edge detection is to decrease number of pixel in object (circle) space searching in Hough transform step.

5.7 Circle Hough Transform

Circle Hough transformation find the central point (x_0,y_0) of each circle object in an image. General formula of circle Hough transform is:

$$(x-x_0)^2+(y-y_0)^2=r^2 \quad (4)$$

Where (x_0,y_0) is center of circle coordinate, r is radius and (x,y) is pixel coordinate on the circle line.

This transformation utilizing the use of accumulators to store the voting results of the candidate (x_0,y_0). Value (x_0,y_0) with the highest voting results or on the threshold value is set as (x_0,y_0) of the detected circle.

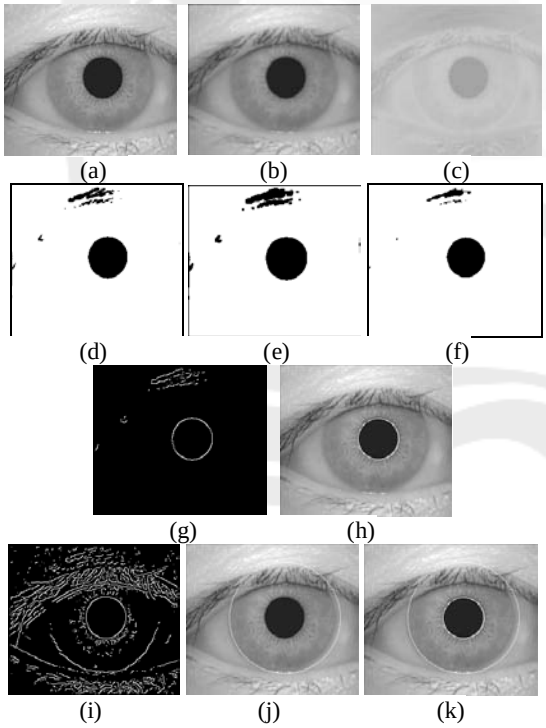


Figure 3. Hough segmentation result, (a) original image, (b) Gaussian image, (c) normalized image, (d) binarized image, (e) dilated image, (f) eroded image, (g) edge image of (g), (h) detected pupil circle, (i) edge image of (b), (j) detected iris circle, (k) segmented pupil and iris circle

To find the object circle in the accumulator need various size of radius (r). The algorithm to find the radius r is as follows:

1. Preparing the accumulator to store the value of r, x_0 and y_0.
2. Determine the upper and lower limit r.
3. For each edge point (x_t,y_t) in the image do step 4.
4. For each value of r in the range of the upper and lower limit r, do step 5.
5. For each value of theta (0-360) do step 6 and 7. Step 5, 6 and 7 is the step description of the circle by setting the edge point (x_t,y_t) as the central point.
6. Calculate the point (x, y) which is as far as r and the angle of θ from (x_t,y_t).
7. Make a vote by adding the value of the accumulator in the coordinates (x, y, r) with 1.

Value of x_0, y_0 and r is taken from the coordinates of the accumulator with the highest voting score.

The results of all steps of Hough segmentation method are shown in Figure 3. Circle Hough transform is used to find pupil circle from the edge image of eroded image (see figure 3 (g,h)) and iris circle from edge image of Gaussian image (see figure 3 (i,j)).

6. DAUGMAN SEGMENTATION

Steps of Daugman segmentation methods are: contrast stretching, image binarization, dilation, erosion, center of area finding using moment, image loading, contrast stretching, Gaussian filtering, pupil circle and iris circle finding using Daugman method respectively. The image binarization, dilation, erosion and Gaussian filtering processes are same with Hough method.

6.1 Contrast Stretching

Contrast stretching increases the contrast value of iris image so the different between pupil and others in image become clearly. The contrast stretching algorithm is as follows:

1. Input contrast factor C
2. Set max = 127 + (C x -1)
3. Set min = 127 – (C x -1)
4. For each pixel in image do the following:
5. Compute new pixel value by formula:
   val = min + ((pixel value/255) x (max – min))
6. If val is greater than 255 then set val = 255
7. If val is less than 0 then set val = 0
8. Set pixel value = val.

6.2 Center of Moment

The purpose of center of moment computation is to find the center point of pupil. This process is performed on eroded image (binary image from erosion process result). The algorithm to computer the center of moment coordinate is:

1. Set $m_{00} = 0$
2. Set $m_{10} = 0$
3. Set $m_{01} = 0$
4. For each pixel in image do the step 5:
5. If pixel value = 0 (pixel is object) then set:
   Store the pixel coordinate as (x,y)
   $m_{00} = m_{00} + 1$
   $m_{10} = m_{10} + x$
   $m_{01} = m_{01} + y$
6. Compute the center of moment coordinate (Mx,My):
   $Mx = m_{10}/m_{00}$

$My = m_{01}/m_{00}$

(Mx,My) coordinate is the center of pupil and this coordinate will be used for finding of Daugman circle.

6.3 Daugman Circle Finding

The algorithm to find the Daugman circle is based on Integro differential Daugman operator. This operator is defined as follows:

$$Max_{(r,x_0,y_0)}\left|G_\sigma(r) * \frac{\partial}{\partial r}\oint_{r,x_0,y_0}\frac{I(x,y)}{2\pi r}ds\right| \quad (5)$$

Where *I(x,y)* represents the eye image, *r* is the radius that will be find, $G_\sigma(r)$ is Gaussian function and s is the circle line that is formed using *r*, x_0, y_0. This operator works by seeking the circle path where the differentiation of summation of pixels intensity with radius r occurs.

This method will be used to find the pupil and iris circle. To find these circle, this algorithm need a center point as input. Based on the point, the Daugman circle can be detected by computing the differentiation of summation of pixels intensity at radius r and radius r+1.

Below is the algorithm to compute the Daugman circle.

1. Input central point, interval value of r from r1 to r2 and max=0
2. For each r value on the interval r1 to r2 do step 3-6:
3. Draw the circle with radius r and compute the summation of their pixels intensity. We called this value as Sr.
4. Draw the circle with radius r+1 and compute the summation of their pixels intensity. We called this value as Srr.
5. Compute: Sd = abs (Sr – Srr)
6. Compare Sd with max. If Sd is greater than max then set max = Sd and store the value of r.
7. The selected r is the radius with maximum Sd.

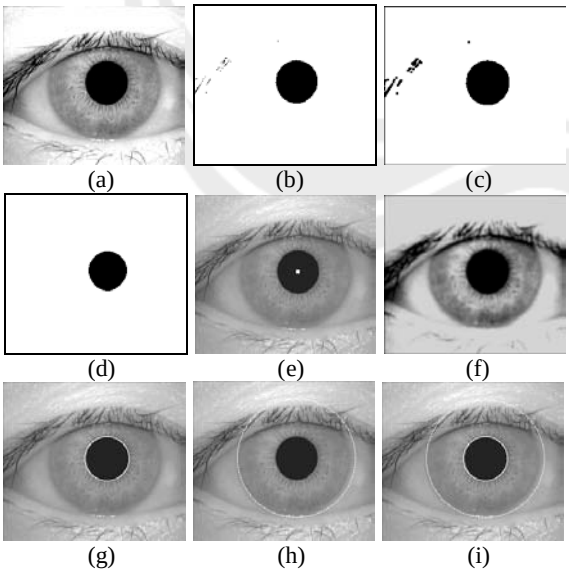


Figure 4. Daugman segmentation result, (a) contrast stretching image of image in figure 3(a), (b) binarized image, (c) dilated image, (d) eroded image, (e) center of moment of image (d), (f) Gaussian image of (a), (g) pupil circle, (h) iris circle, (i) segmented pupil and iris circle.

7. EYELID DETECTION

Eyelids and eyelashes that cover the iris are two types noise in iris image. Eyelid detection is done after segmentation process while eyelash detection is performed on normalized image. The algorithm to detect the eyelid is as follows:

1. Initial value : Cmt = 0, Cmb = 0.
2. Perform edge detection to find horizontal edge of image.
3. For each line of pixels that lie between the top of pupil and iris section do the following:
4. Count the number of edge pixels on the line. We called this value as Ct. Compare the value Ct with Cmt. If Ct is larger than Cmt then set Cmt = Ct, and store the ordinate position (Yt) of the line.
5. For each line of pixels that lie between the bottom of the pupil and the iris section do the following:
6. Count the number of edge pixels the line We called this valus as Cb. Compare the value Cb with Cmb, if Cb is larger than Cmb then set Cmb=Cb, and store the ordinate position (Yb) of the line.
7. Fill all lines above the ordinate Yt if the Cmt is greater than the threshold value.
8. Fill all lines below the ordinate Yb if the Cmb is greater than the threshold value.

Figure 5 show the example of eyelid detection result.

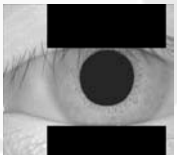


Figure 5. Eyelid Detection Results.

8. NORMALIZATION

The normalization transforms the iris image after eyelid detection in Cartesian coordinate to the polar coordinate system.

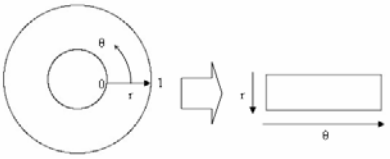


Figure 6. Cartesian to polar transformation

This process transform each iris image pixel coordinate (x,y) to the polar coordinate (r,θ) where *r* (radius) value between 0-1 and θ value between 0 - 2π. This transformation can be illustrated as follows:

$$I(x(r,\theta), y(r,\theta)) \longrightarrow I(r,\theta) \quad (6)$$

With criterion:

$$x(r,\theta) = (1-r)x_p(\theta) + rx_i(\theta) \quad (7)$$

$$y(r,\theta) = (1-r)y_p(\theta) + ry_i(\theta) \quad (8)$$

Where $I(x,y)$ represents iris image before transformation with Cartesian coordinate (x,y) and (r,θ) represents polar coordinate after normalization. $(x_p , y_p,)$ and (x_i , y_i) represents pupil and iris boundary coordinate respectively at path of θ [9].

Normalization process is done by applying the transformation of coordinates to polar coordinates Cartesian. Here is the normalization process is:

1. Form a new image with sized 360 in width and R Iris - R pupils in length.
2. For each point (θ, r) in the new image, calculate coordinates (x, y) in the original image.
3. Use the intensity at point (x, y) at the original image as the pixel value at the point (θ, r) in the polar image.

Figure 7. Normalization Results.

9. EYELASH DETECTION

Eyelash detection process is binarization process. Eyelash will be noted as black pixels after binarization. Figure 8 show the result of eyelash detection.

Figure 8. Eyelash Detection Results.

10. EXPERIMENT AND RESULT

The performances of two methods are measured by using CASIA databases. The CASIA databases are split into several groups of databases. Because the number of different subject and the number of samples for each subject are different for each CASIA database then the number of groups for each database also different.

Table 1, 2, 3 and 4 show the experiment results using databases that vary in size with 25, 50, 75 and 100 different subjects respectively with different number of samples for each subject. The success rate is shown in the tables are computed based on subjective scoring by users.

Table 1. The success rate of Hough and Daugman segmentation method on database contained 25 People

Database	Index database	Hough	Daugman
CASIA 1	Database 1	94.86%	92.57%
CASIA 1	Database 2	93.71%	88.00%
CASIA 1	Database 3	96.57%	90.86%
CASIA 1	Database 4	93.71%	91.43%
CASIA 2	Database 1	52.00%	26.75%
CASIA 2	Database 2	54.50%	25.75%
CASIA 3	Database 1	93.73%	35.69%
CASIA 3	Database 2	91.88%	49.57%
CASIA 3	Database 3	92.19%	45.72%
CASIA 3	Database 4	93.97%	56.03%
CASIA 3	Database 5	90.66%	43.96%
CASIA 3	Database 6	94.83%	35.26%
CASIA 3	Database 7	83.86%	38.60%
CASIA 3	Database 8	92.92%	49.53%
CASIA 3	Database 9	91.43%	44.29%
CASIA 3	Database 10	81.21%	42.28%

Table 1 show the comparison of the success rate of two methods on databases that contained 25 different subjects. The CASIA 1 database is divided into 4 groups of database (e.g.: database 1, database 2, database 3 and database 4) are shown at index database column. The CASIA 2 database is split into 2 groups of database (e.g.: database 1 and database 2), and the CASIA 3 database is divided into 10 groups (e.g.: database 1 to database 10).

The number of group is different for each CASIA database because the number of subjects in each CASIA database different. The Hough and Daugman methods are performed to each group and the success rate is computed by users using subjective scoring witch match or not match criterion. The number of match and not match score are computed, and the success rate is defined as follows:

Success rate = (total of match score / total of samples) x 100%.

Table 2. The success rate of Hough and Daugman segmentation method on database contained 50 People

Database	Index Database	Hough	Daugman
CASIA 1	Database 1	94.29%	90.29%
CASIA 1	Database 2	95.14%	91.14%
CASIA 2	Database 1	53.25%	26.25%
CASIA 3	Database 1	92.84%	42.33%
CASIA 3	Database 2	93.19%	51.54%
CASIA 3	Database 3	92.64%	39.83%
CASIA 3	Database 4	87.73%	43.26%
CASIA 3	Database 5	83.21%	25.91%

In table 3 and 4 is shown that the performances of the two methods are not measured in CASIA 2 database because the number of subject in that database is less than 75 different subjects.

Table 3. The success rate of Hough and Daugman segmentation method on database contained 75 People

Database	Index Database	Hough	Daugman
CASIA 1	Database 1	95.05%	90.48%
CASIA 3	Database 1	92.61%	43.54%
CASIA 3	Database 2	93.08%	45.24%
CASIA 3	Database 3	88.83%	43.56%

Table 4. The success rate of Hough and Daugman segmentation method on database contained 100 People

Database	Index Database	Hough	Daugman
CASIA 1	Database 1	94.71%	90.71%
CASIA 3	Database 1	93.04%	47.47%
CASIA 3	Database 2	92.72%	40.86%
CASIA 3	Database 3	90.26%	42.80%

In all of testing, the performance of Hough method is more accurate than Daugman method.

Comparison of time complexity of Daugman and Hough method are shown di Table 5.

Table 5 Time complexity of Hough and Daugman method

Database	Daugman (s)	Hough (s)
CASIA 1	± 0.67	± 2.16
CASIA 2	± 0.62	± 1.67
CASIA 3	± 1.16	± 1.61

11. CONCLUSION

Two methods of iris segmentation, Hough and Daugman method, has been tested in this paper. In all of experiment results by using three CASIA databases, the Hough method has better accuracy than Daugman method, but in time complexity, Daugman method more efficient relatively than Hough method.

12. REFERENCES

[1] Arvacheh, Ehsan M. 2006. Tesis: A Study of Segmentation and Normalization for Iris Recognition Systems. Kanada : The University of Waterloo.

[2] Chinese Academy of Sciences Institute of Automation. 2006. CASIA-IrisV3. http:\\www.cbsr.ia.ac.cn/IrisDatabase.htm.

[3] Darma Putra, IKG. 2006. Disertasi : Metode Fraktal untuk Sistem Pengenalan Biometrika Telapak Tangan. Yogyakarta : Universitas Gajah Mada.

[4] Darma Putra, IKG. 2009. Sistem Biometrika Teori dan Aplikasi. Andi Offset : Jogjakarta

[5] Darma Putra, IKG. 2010. Pengolahan Citra Digital. Andi Offset : Jogjakarta

[6] Daugman, John. 2002. How Iris Recognition Works. Proceeding of 2002 International Conference on Image Processing. Vol.1.

[7] Jain, Anil K., Bolle, Ruud., dan Pankanti, Sharath. 1999. Introduction to Biometrics. BIOMETRICS Personal Identification in Networked Society. London : Kluwer *Academic Publisher*.

[8] Lipinski, Bryan. 2004. Iris Recognition : Detecting The Iris. http://cnx.org/content/m12489/1.3/

[9] Masek, Libor. 2003. Recognition of Human Iris Patterns for Biometrics Identification. The University of Western Australia.

[10] Robichaux, Paul., Khabashesku, Dmitry. Iris Recognition Results and Conclutions. http://cnx.org/content/m12495/1.2/

[11] US Commercial Service. 2005. Biometrics an Emerging Technology. Market Report May 2005 – US Commercial Service Australia

NEATS: A New Method for Edge Detection

Maria Yunike
Informatics Engineering Department
Universitas Atma Jaya Yogyakarta
nicke_mail@yahoo.com

Suyoto
Informatics Engineering Department
Universitas Atma Jaya Yogyakarta
suyoto@mai.uajy.ac.id

ABSTRACT
This paper presents a new method for edge detection called NEATS. NEATS is an innovative edge detection method for digital image processing. Edge detection is used to identify an image by analyzing the image boundaries to determine the edge of the image. This new method uses a combination matrix from Elisabeth Method, Thomas Method, Silny Method and Adhi Method. We analyze the advantages and disadvantages of these methods to generate new matrix that can perform edge detection better. We applied this new method into various types of Indonesian Batik motives. When this new method implemented, it delivered different result for each batik motives. The experimental results indicate that the composition of the matrix which used for edge detection is influenced by the image pattern that will be tested. An image that tested by using the appropriate matrix will generate the optimal edge detection result.

Keywords
NEATS, Edge Detection, digital image processing, Indonesian Batik motives.

1.INTRODUCTION
Image processing becomes one of the most challenging problems in Computer Vision [1]. The purpose of image processing is to improve the image quality to be easily interpreted by humans and machines. Image processing techniques used to transform the image into another image. Image processing operations can be classified in several types of image quality improvement, such as: image restoration, image segmentation, image analysis and image reconstruction [2]. Image quality improvement is used to produce new images with better quality than the original image. Image analysis is used to identify the parameters associated with the characteristics of the object in the image, to further these parameters used in interpreting the image [2]. Key factor in extracting features of objects in the image is the ability to detect the edge of the image [2]. An object can be identified by the human eye really well. The human eye has a very good ability to identify an image, including contrast and color threshold [3]. However, the machine requires considerable effort to identify an image. Computer uses edge detection method to identify a digital image.

Edge detection is a very important feature-extraction method that has been widely used in many computer vision and image processing applications. The basic idea of most available edge detectors is to locate some local object-boundary information in an image by thresholding the pixel-intensity variation map [4]. In edge detection, we examine the image's pixel by comparing the changes with its neighbors [3]. This approach will find the edge threshold image based on the color contrast. The problem happens when the contrast on the image's threshold is not clear. Poor color differences on the edge of the image reduces the accuracy of the edge detection process [3]. Therefore, we need a better method to determine where the edges are.

Although many edge-detection evaluation methods have been developed, this is still a challenging and unsolved problem. The major challenge comes from the difficulty in choosing an appropriate performance measure of the edge-detection results [4]. We explore several methods for edge-detection, as detailed in the next section. This includes Elisabeth Method, Thomas Method, Silny Method and Adhi Method. To determine the correctness of each method, we compare the results of edge detection by using those methods. Then, we analyze the advantages and disadvantages of those methods to generate new matrix that can perform edge detection better. In this study, we applied the new method into various types of Indonesian batik motives such as Batik Parang, Sidomukti, and Semen. The purpose of this research is to apply the NEATS method into Indonesian batik motives.

2. EDGE DETECTION METHOD
Boundary edge of an image is often referred to as the edge. These boundaries are discovered by following a path of rapid change in image intensity [3]. Edge information for a particular pixel is obtained by exploring the brightness of pixels in its neighborhood. If all of the pixels in the neighborhood have almost the same brightness, then there is probably no edge at that point. However, if some of the neighbors are much brighter than the others, then there is a probably an edge at that point. Measuring the relative brightness of pixels in a neighborhood is mathematically analogous to calculating the derivative of brightness. Brightness values are discrete, not continuous, so we approximate the derivative function. Different edge detection methods (Prewitt, Laplacian, Kirsch, Sobel etc.) use different discrete approximations of the derivative function. They look for places in the image where the intensity changes rapidly by locating places where the first derivative of the intensity is larger in magnitude than some threshold, or finding places where the second derivative of the intensity has a zero crossing [5].

The basic edge-detection operator is a matrix area gradient operation that determines the level of variance between different pixels [6]. The edge-detection operator is calculated by forming a matrix centered on a pixel chosen as the center of the matrix area. If the value of this matrix area is above a given threshold, then the middle pixel is classified as an edge [6]. There are many types of edge detection algorithms such as gradient edge-detectors, Laplacian of Gaussian (LOG), zero crossing, and Gaussian edge-detectors [7]. Sobel, Prewitt and Roberts method are the example of gradient edge-detectors method. These methods are looking for maximum and minimum first derivative of an image. Marr and Hildreth found Laplacian of Gaussian (log) approach that combines the Gaussian filtering with zero crossing for the second derivative of the image [8]. This method

detects the location of the edge, especially on the steep edge of the image [2]. At the steep edge of image, the second derivative value has zero crossing. It is the point where there is a change of second derivative values. On the sloping edge, there is no zero crossing [2]. The zero crossing method searches for the zero crossing in the second derivative of the image is like Laplacian of Gaussian, however it does not use Gaussian filtering [9]. Gaussian edge-detectors detect the image edges symmetrically along the edge and reduce noise by smoothing the image [3].

In order to detect edges of an image, we can use an edge detector program that using convolution filters. In each case, we apply a horizontal version of the filter to one bitmap, a vertical version another, then we use the formula pixel (1) to merge them together.

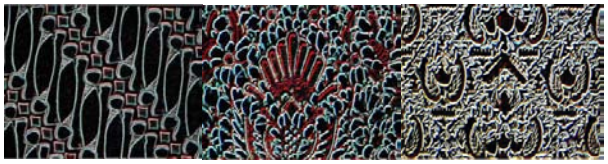
$$\text{Formula pixel} = sqrt(pixel1 * pixel1 + pixel2 * pixel2) \quad (1)$$

The convolution masks equally with the matrix that used for detection. These filters perform the horizontal edge detect, rotating them 90 degrees gives us the vertical, and then the merge takes place. Edge detection filters work essentially by looking for contrast in an image. This can be done by many different ways. The convolution filters do it by applying a negative weight on one edge, and a positive on the other. This has the net effect of trending towards zero if the values are the same, and trending upwards as contrast exists. By using this program, we can set a matrix that is used as a filter and then perform edge detection on an image to produce a new image that showed the result of edge detection.

Many studies have been developed to find the edge detection algorithm that produces better results. Several new methods are developed based on edge detection algorithms that already exist. These new method are Elisabeth method, Thomas method, Adhi Method and Silny Method. Elisabeth method developed gradient method combined edge-detectors of Prewitt and Sobel by using multiplication by the identity matrix [10]. In Elisabeth method, the matrix x derived from the matrix x from Prewitt Method. Matrix y derived from matrix x from Sobel Method [10]. Matrix that used in Elizabeth Method shows in (2).

$$Ex = \begin{bmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{bmatrix} \qquad Ey = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} \quad (2)$$

The implementation of Elizabeth method shows that the straight line of the image can be detected clearly. However, if the image contains many curve lines, the result of edge detection seems blurred. This happens because the matrix x in Elizabeth method identified vertical straight lines while the matrix y identified horizontal straight lines. So the curve lines can not be detected accurately especially when the color is quite same [10]. By using this method, the vertical lines seems thicker on the right side because there are positive value on the right side of matrix x. The horizontal lines seem thicker on the bottom side because of the positive value on the bottom side of matrix y. The result of Elizabeth method implementation on Batik Image can be seen in Figure 1.



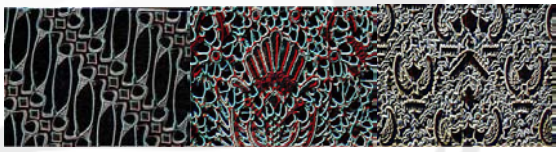
(a) (b) (c)

Figure 1. Edge Detection on Indonesian Batik Motives: (a) Batik Parang, (b) Batik Sidomukti, (c) Batik Semen tested by Elizabeth Method

Thomas method is combination between Prewitt and Canny method. In Thomas method, the matrix x derived from Prewitt Method that has been multiplied by the identity matrix and matrix y derived from Canny Method [11]. Matrix that used in Thomas Method shows in (3).

$$Ex = \begin{bmatrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{bmatrix} \qquad Ey = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \quad (3)$$

By using Thomas Method, we can identify straight lines clearly. But, curve lines still looks blurred. Positive value on the left side of matrix x shows the thicker line on the left side of the line. Positive value on upper side of matrix y shows the thicker line on the upper side of the line. Thomas method has been tested on Batik Image. The result of Thomas Method implementation can be seen in Figure 2.



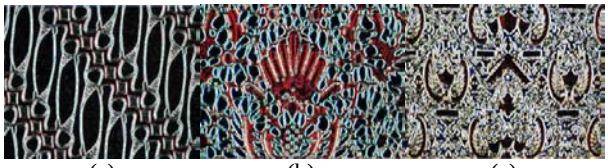
(a) (b) (c)

Figure 2. Edge Detection on Indonesian Batik Motives: (a) Batik Parang, (b) Batik Sidomukti, (c) Batik Semen tested by Thomas Method

Silny method is combination between Canny and Sobel method. In Silny method, the matrix x derived from the matrix x from Canny Method and matrix y derived from matrix y from Sobel Method [12]. Matrix that used in Silny Method shows in (4).

$$Ex = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \qquad Ey = \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix} \quad (4)$$

By using Silny method, straight lines and curved lines on the boundaries of the object can be seen clearly. However, the edge details can not be identified. Implementation of Silny method on Batik Image can be seen in Figure 3.



(a) (b) (c)

Figure 3. Edge Detection on Indonesian Batik Motives: (a) Batik Parang, (b) Batik Sidomukti, (c) Batik Semen tested by Silny Method

Adhi method is developed from Quick Mess method. In this method, the Quick Mess matrix modified by moving the composition of the matrix value so that formed the PLUS sign on the center of matrix [13]. This form makes the horizontal and vertical lines of the object can be identified well. Then, the value of matrix increased three times higher [13]. By increasing the value of Quick Mess, edge can be identified more accurately. Matrix that used in Adhi Method shows in (5).

$$Ex = \begin{bmatrix} 0 & -3 & 0 \\ -3 & 12 & -3 \\ 0 & -3 & 0 \end{bmatrix} \quad Ey = \begin{bmatrix} 0 & -3 & 0 \\ -3 & 12 & -3 \\ 0 & -3 & 0 \end{bmatrix} \qquad (5)$$

By using Adhi method, curved lines and straight lines can be identified very details but the lines was very tiny. In addition, there is also a lot of noise caused by the highly accurate detection on the image. The result of Adhi Method implementation tested on Batik Image can be seen in Figure 4.

(a) (b) (c)

Figure 4. Edge Detection on Indonesian Batik Motives: (a) Batik Parang, (b) Batik Sidomukti, (c) Batik Semen tested by Adhi Method

3. NEATS METHOD

Based on above explanation, we know that each matrix has its own advantages and disadvantages. A method that applied to different images will produce different quality. We found that these differences depend on the dominance pattern contained on the image. Basically, the image is usually composed of vertical lines, horizontal lines and curve lines. Each pattern requires different edge detection process. As we know, Elizabeth and Thomas Method produces good result of edge detection for straight lines but not too good to detect the curve lines. While Adhi Method produces good edge detection for curve lines with details, although the lines seem very tiny.

We found that a straight line can be detected by either if the value of the matrix composed straight. So, if we want to strengthen the edge detection on the vertical line, we have to use a matrix that has value which composed vertically on the border of matrix. While, if we want to strengthen the edge detection on the horizontal line, we have to use a matrix which value composed horizontally on the border of matrix.

However, some images contain more curve lines than the straight lines. Therefore, we need a method that can detect curve lines very well. Adhi method is good enough to detect the curve lines, although edge detection on the straight lines is very tiny. By using Adhi Method, the boundaries of the image can be identified accurately. This is because the large value on the center of the matrix causes the dark color and the negative value on border of matrix causes the bright color so that the edge of the image will be seen.

We found that in order to detect an accurate edge on the boundaries of the image, we have to combine the matrix. Matrix of Adhi Method can be used as the main matrix (Ex) to obtain images detailed. Then, we use matrix from Elizabeth and Thomas Method to identify the straight line by using it as matrix y (Ey). In fact, there are many images that almost consist of vertical straight line and there are other images that almost consist of horizontal straight line. Equation (6) is appropriate matrix for identifying images that have vertical straight line pattern. Equation (7) is appropriate matrix for identifying images that have horizontal straight line pattern.

$$Ex = \begin{bmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{bmatrix} \qquad (6)$$

$$Ey = \begin{bmatrix} -1 & -1 & -1 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix} \qquad (7)$$

Therefore, we developed two types of matrix models in this method: NEATS 1 and NEATS 2. sssOn NEATS1 we use matrix from Adhi Method as the main matrix (Ex) and use matrix in (6) as its matrix y (Ey). Matrix that will be used in NEATS1 showed in (8). By using NEATS1, we expect to detect the edge of the image that almost contains the vertical straight line pattern. On NEATS2 we use matrix from Adhi Method as the main matrix (Ex) and use matrix in (7) as its matrix y (Ey). Matrix that will be used in NEATS2 showed in (9). By using NEATS2, we expect to detect the edge of the image that almost contains the horizontal straight line pattern. Both matrix models will be tested to measure the accuracy of edge detection on many various pattern of Indonesian Batik motive.

$$Ex = \begin{bmatrix} 0 & -3 & 0 \\ -3 & 12 & -3 \\ 0 & -3 & 0 \end{bmatrix} \quad Ey = \begin{bmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{bmatrix} \qquad (8)$$

$$Ex = \begin{bmatrix} 0 & -3 & 0 \\ -3 & 12 & -3 \\ 0 & -3 & 0 \end{bmatrix} \quad Ey = \begin{bmatrix} -1 & -1 & -1 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix} \qquad (9)$$

4. EXPERIMENTAL RESULT

When implemented on Indonesian Batik motives, NEATS delivered various result for each Batik motives. We determine three types of general batik motif, such as batik with horizontal straight line dominance, batik with vertical straight line dominance and batik with curve line dominance.

We use three Batik motives to represent each type of Indonesian batik motive for NEATS implementation. There are Batik Parang, Sidomukti and Semen. Each batik motive has own characteristic. Batik Parang has many straight lines. Some of the curve lines are very tiny and adjacent but not sharp enough.

Batik Sidomukti has many complicated curve lines. Batik Semen has balance curve lines and straight lines. These Indonesian Batik motives show in Figure 5.

(a) (b) (c)

Figure 5. Indonesian Batik Motives: (a) Batik Parang, (b) Batik Sidomukti, (c) Batik Semen for NEATS Implementation

Then, we applied NEATS1 and NEATS2 on those images. This is the result of NEATS method on Indonesia Batik Motives:

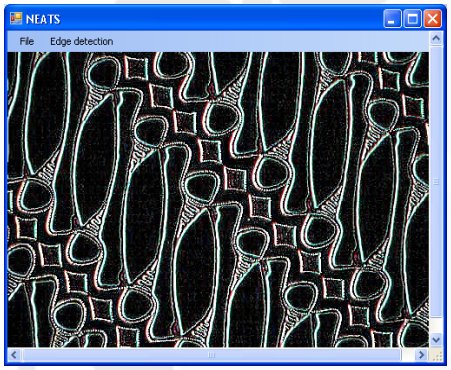


Figure 6. NEATS1 implemented on Batik Parang

Figure 7. NEATS2 implemented on Batik Parang

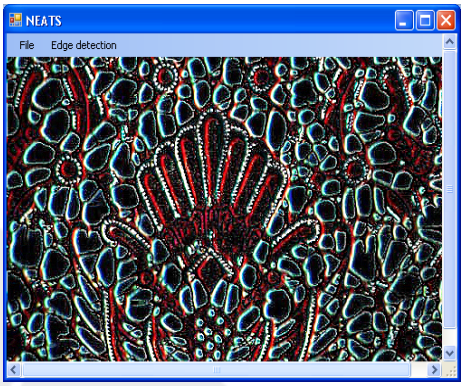


Figure 8. NEATS1 implemented on Batik Sidomukti

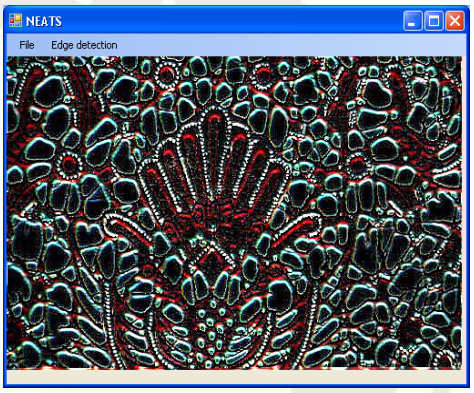


Figure 9. NEATS2 implemented on Batik Sidomukti

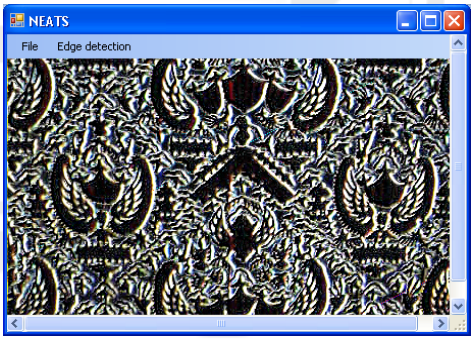


Figure 10. NEATS1 implemented on Batik Semen

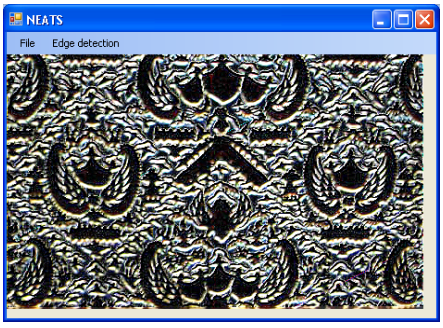


Figure 11. NEATS2 implemented on Batik Semen

After NEATS had been implemented, we held a survey on 100 peoples to identify the results. First, we compare the results of edge detection on each batik motive with Elisabeth Method, Thomas Method, Silny Method, Adhi Method and NEATS Method. We provide six levels of rank on each batik motive to determine the quality of edge detection result from the highest to the lowest. The result can be seen in Table 1.

Table 1. Comparison of Edge Detection Result on Indonesian Batik Motives

	Batik Parang	Batik Sidomukti	Batik Semen
Elisabeth Method	5	5	6
Thomas Method	4	6	5
Silny Method	6	4	4
Adhi Method	3	3	3
NEATS 1	1	1	2
NEATS 2	2	2	1

Based on the Table 1, we know that the edge detection with NEATS 1 produces the best results for Batik Parang and Batik Sidomukti. Edge Detection using NEATS 2 is most suitable for Batik Semen. This identification is based on the overall quality of edge detection result on each Batik Motive.

Edge detection with NEATS 1 produces the best result on Batik Parang that has a lot of vertical lines. NEATS 2 actually good enough implemented for Batik Parang, but NEATS 2 are more suitable for an image that has many horizontal lines like Batik Semen. Edge detection using Adhi Method on Batik Parang has actually been quite good and detailed, but the margins are too thin and not clear. Elizabeth Method and Thomas Method also produce good edge detection result on Batik Parang. However, these methods can not detect the edges on the images in detail.

On Batik Sidomukti that has a lot of curved lines, NEATS 1 also produces the best result. Actually NEATS 1 and NEATS 2 suitable for the edge detection on images that have many curved lines. Adhi Method has actually good enough for edge detection on Batik Sidomukti. It can produces result with great detail but unfortunately the lines were too thin. Edge detection on Batik Sidomukti with Silny Method is also quite good but the results are not detailed. Elizabeth Method and Thomas Method are not suitable used for edge detection on Batik Sidomukti because both method can not identify overall lines clearly.

On Batik Semen which has many horizontal lines, NEATS 2 provides the best edge detection result. NEATS 1 also actually produces good result for edge detection on Batik Semen, but NEATS 1 is more suitable for an image that has a lot of vertical lines such as Batik Parang. In fact,the previous methods also good enough implemented on Batik Semen. However, edge detection results with previous methods have many disadvantages. Edge detection result using Adhi Method on Batik Semen already detailed but unclear. Edge detection result using Silny Method was pretty clear but not detailed enough. Elizabeth and Thomas Method method is not suitable for edge detection on Batik Semen because these methods can not detect horizontal lines cleary.

Second, we try to identify the accurate of edge detection on Indonesian Batik motives by NEATS1 and NEATS2. We also give five range levels to identify each identification item of the edge detection result. The result can be seen in Table 2.

Table 2. NEATS on Indonesian Batik Motives

	Vertical Stright Lines	Horizontal Straight Lines	Curve Lines	Noise	Details
NEATS 1 on Batik Parang	5	4	5	2	5
NEATS 2 on Batik Parang	4	5	5	3	5
NEATS 1 on Batik Sidomukti	5	4	5	2	5
NEATS 2 on Batik Sidomukti	4	5	5	3	5
NEATS 1 on Batik Semen	5	4	5	2	5
NEATS 2 on Batik Semen	4	5	5	2	5

Based on the Table 2, we know that if NEATS1 applied on Batik Parang, the vertical straight lines will be seen clearly with low noise. The curve lines also can be identified very well in details. However, horizontal straight lines cannot be identified horizontal as well as vertical straight line. Horizontal straight lines seem blurred. If NEATS2 applied on Batik Parang, the horizontal straight lines will be seen clearly. But, the vertical straight lines seem not too clearly with more noise. The curve lines can be identified very well in details.

When NEATS1 and NEATS2 applied on Batik Sidomukti, the curve lines can be identified very clearly. But, when using NEATS1, vertical straight lines can be identified clearly while horizontal straight lines seem blurred. When using NEATS2, horizontal straight lines can be identified clearly while vertical straight lines seem blurred.

Batik Semen has many curve lines and horizontal straight lines pattern. When NEATS1 and NEATS2 applied on it, we found that curves lines can be identified very clearly. By using NEATS1, vertical straight lines can be seen clearly but horizontal straight lines seem blurred with low noise. By using NEATS2, horizontal straight lines can be seen clearly but vertical straight lines seem blurred with high noise.

5. CONCLUSION

Based on the experimental result, we found that image edge detection is influenced by the matrix. The matrix selection is based on the type of image that will be tested. So the composition value should be placed exactly according to the dominant pattern on the image that will be tested. By NEATS method, we can perform accurate edge detection on digital images that have different lines domination. NEATS method is suitable for edge detection on Indonesian Batik Motives that has many types of patterns and different structure. By using NEATS1, we can perform accurate edge detection on images that have lots of curve and vertical lines. By using NEATS2, we

can perform accurate edge detection on images that have lots of curve and horizontal lines. An image that tested by using the appropriate matrix will generate the optimal edge detection result.

6. REFERENCES

[1] Roman Louban. *Image Processing of Edge and Surface Defect*. Springer Series in Materials Science , Vol. 123. June 2009

[2] Dewi Agushinta R, Alina Diyanti. Perbandingan Kinerja Metode Deteksi tepi pada Citra Wajah. In Jurnal Universitas Gunadarma edisi Nomor 2 Volume 1, Mei 2008.

[3] Evelyn Brannock, Michael Weeks. Edge Detection Using Wavelets. Algorithms. *In Proceedings of the 44th annual ACMSoutheast. Regional Conference (ACMSE 2006)*, pages 649–654, Mar 2006.

[4] Song Wang, Feng Ge, and Tiecheng Liu. *Evaluating Edge Detection through Boundary Detection*. Department of Computer Science and Engineering, University of South Carolina, Columbia, USA. June 2005.

[5] The MathWorks, Inc. Image Processing Toolbox User's Guide. The MathWorks, Inc, Natick, MA, 2004.

[6] Hong Shan Neoh and Asher Hazanchuk. Adaptive Edge Detection for Real-Time Video Processing using FPGAs. Department of Computer Science and Engineering, University of South Carolina, Columbia, USA. June 2005.

[7] M. Shari, M. Fathy, and M. T. Mahmoudi. A classified and comparative study of edge-detection algorithms. In *Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC.02)*, pages 117-120, April 2002.

[8] D. Marr and E. Hildreth. Theory of edge-detection. In *Proceedings of the Royal Society of London. Series B*, volume 207, pages 187-217, 1980.

[9] S. Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Pattern Analysis and Machine Intelligence*, 11(7):674-693, 1989.

[10] Elisabeth Dhenys. *New Edge Detection using Elisabeth Method: Case Study Javanese Batik* (Unpublished). Universitas Atma Jaya Yogyakarta, December 2009.

[11] Thomas Adi Purnomo Shidi, *New Edge Detection Method for Indonesian Batik* (Unpublished). Universitas Atma Jaya Yogyakarta, December 2009.

[12] Silvia. *Silny approach to Edge Detection for Central Borneo Batik* (Unpublished). Universitas Atma Jaya Yogyakarta, December 2009.

[13] Adhi Pranoto. *Development Edge Detection using Pranoto Method, Case Study : Batik Sidomukti Motive* (Unpublished). Universitas Atma Jaya Yogyakarta, December 2009.

Online Facial Caricature Generator

Rudy Adipranata
Informatics Department
Petra Chistian University
Siwalankerto 121-131
Surabaya, Indonesia
62-31-8439040
rudya@petra.ac.id

Stephanus Surya Jaya
Informatics Department
Petra Chistian University
Siwalankerto 121-131
Surabaya, Indonesia
62-31-8439040
east_62687@hotmail.com

Kartika Gunadi
Informatics Department
Petra Chistian University
Siwalankerto 121-131
Surabaya, Indonesia
62-31-8439040
kgunadi@petra.ac.id

ABSTRACT
Facial caricature is a portrait of a person's face where the unique parts of the face is exaggerated. Nowadays, there are many website that provide caricature generation services. However, the resulting caricature was done by caricaturist. Based on that, a website that able to generate caricature from human face photograph is made in this research. The process is defined as follows: the image containing the face is uploaded to the website, facial feature extraction process to extract facial feature points from the image, shape exaggeration process to exaggerate unique facial features and caricature generation process that is defined as an image warping process to a prepared caricature. The experimental shows the resulting caricature has exploited unique facial features.

Keywords
Caricature generator, facial features, shape exaggeration

1. INTRODUCTION
Caricatures usually made by experienced caricaturist. There are many websites that offers caricature generation services by these caricaturist. In this research, an online caricature generator application that replace the role of caricaturist to generate a caricature from a face photograph is developed. The user of this application only need to upload the photograph, locate the face and determine several oher parameters that is needed for the caricature generation process. There are some research about caricature generation [1,2,6]. The most important process in this caricature generation process is the facial feature extraction process and the shape exaggeration process to exaggerate the facial features.

2. CARICATURE GENERATOR
Similar to how caricaturist works by exaggerating unique facial features, there are two main processes in this caricature generation process, the first is the facial feature extraction process and the second is shape exaggeration process for unique facial features. The caricature generation process is ended by image warping process with a caricaturist work as the base image and the exaggerated facial feature points as destination point. This research used Active Shape Model [3,4] to find the facial features. The model that is used in this facial feature extraction is based on face model definition proposed by Chiang, et.al [2] based on MPEG-4 face definition parameters [7]. After finding the facial features, the shape exaggeration process is implemented using method that is proposed by Chiang, et.al [2]. The end of the caricature generation process is image warping [5] with a caricaturist work as the base.

3. FACE MODEL DEFINITION
Chiang, et.al [2] explained that to control the shape and appearance of each facial feature, we should define a set of 119 nodes based on the MPEG-4 face definition parameters and face animation parameters. These nodes are categorized into 8 groups: face contour, left-right eyebrow, left-right eye, nose, upper-lower lip. The groups are related to each other by a predefined hierarchy shown in Table 1.

Table 1. Structure and Hierarchy of Face Component [2]

Group ID	*Region*	*Rank*	*NodeCount*
G1	Face Contour	1	19
G2	Left Eye	2	22
G3	Right Eye	2	22
G4	Nose	2	22
G5	Left Eyebrow	3	8
G6	Right Eyebrow	3	8
G7	Upper Lip	3	10
G8	Lower Lip	3	8

Figure 1 consists of an example of face mesh and eight groups of face component with their corresponding master node.

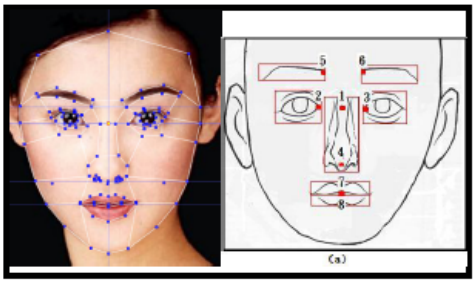


Figure 1. (a) *Face Mesh*. (b) The Eight Groups and Its Corresponding Master Nodes [2]

Each group contains three different types of nodes: master node, calibration node, and slave node. The function of master node is to define the position of the specific group. All other nodes in the same group move accordingly with the master node. The calibration node serves as the reference for shape measurement, normalization and deformation. All nodes that are neither master nodes nor calibration nodes are slave nodes [2]. A quantitative

analysis is also performed in this stage to obtain important statistics of face components. From each sample photograph, each node is marked manually and saved to obtain the statistics and calculate the average face model.

4. ACTIVE SHAPE MODEL

Active Shape Model (ASM) is a method where a model iteratively adapt to refine estimates of the pose, scale and shape of models of image objects [3]. The method uses flexible models derived from sets of training examples.

Given a rough starting approximation, an instance of a model can be fit to an image. By choosing a set of shape parameters b for the model, we define the shape of the object in an object-centred co-ordinate frame. We can create an instance X of the model in the image frame by defining the position, orientation and scale, using [4]:

$$X = M(s,\theta)[x] + X_c \quad (1)$$

where:

- $X_c = (X_c, Y_c, ..., X_c, Y_c)^T$
- $M(s,\theta)[.]$ rotation θ and scale s
- (X_c, Y_c) center position of the model

To summarise, Active Shape Model works as follows [4]:

1. Examine a region of the image around each point find the best nearby match for the point
2. Update pose and shape parameter (X_t, Y_t, s, θ, b) to best fit the new found points
3. Constraint shape parameter (b) to ensure plausible shape. (Example: $|b_i| < 3\sqrt{\lambda_i}$)
4. Repeat until convergence (convergence is reached when there is no significant change between each iteration).

In practice, the search is done along profiles normal to the model boundary through each model point. If the model boundary is expected to correspond to an edge, the strongest edge including orientation if known, can simply be located along the profile. The position of this gives the new suggested location for the model point.

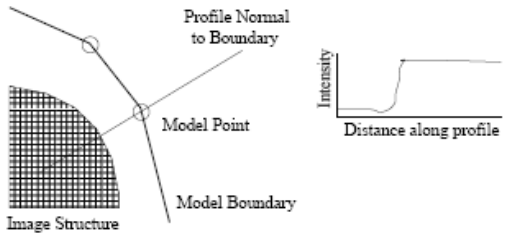


Figure 2. At Each Model Point Sample Along a Profile Normal to The Boundary [3]

However, model points are not always placed on the strongest edge in the locality. They may represent a weaker secondary edge or some other image structure. The best approach is to learn from the training set what to look for in the target image.

4.1 Multi Resolution Active Shape Model

To improve the eficiency and robustness of the algorithm, it is implement in a multi-resolution framework. This involves first searching for the object in a coarse image, then refining the location in a series of finer resolution images [3].

For each training and test image, a gaussian image pyramid is built. The base image (level 0) is the original image. The next image (level 1) is formed by smoothing the original then subsampling to obtain an image with half the number of pixels in each dimension. Subsequent levels are formed by further smoothing and sub-sampling [3].

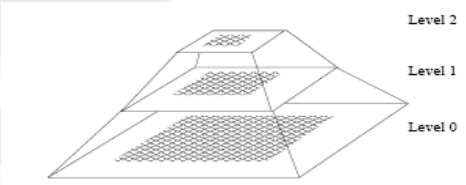


Figure 3. A Gaussian Image Pyramid is Formed by Repeated Smoothing and Sub-sampling [3]

Algorithm for Multi Resolution Active Shape Model [3]:

1. *Set* $L = L_{max}$
2. *While* $L \geq 0$
   a. Compute model point positions in image at level *L*.
   b. Search at each points on profle either side each current point
   c. Update pose and shape parameter (*Xt, Yt, s, θ, b*) to best fit the new found points
   d. Return to 2a if convergence is not reached in level *L* or max iterations have been applied at this resolution.
   e. if $L > 0$ then $L = L - 1$
3. Final result is given by the parameters after convergence at level 0.

5. SYSTEM DESIGN

The developed system is consisted of two main parts. The first is the website which is the main part where the face preparation and caricature generation process is implemented. The second part consists of several programs to mark the face samples and to obtain the statistics that is necessary for facial feature extraction and shape exaggeration process. The overall system design for main part is shown in Figure 4.

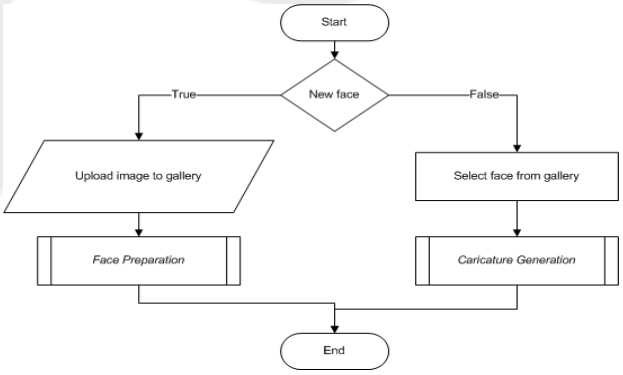


Figure 4. System Design

First, the image containing frontal face photograph is uploaded to the website. The image can be uploaded from user's computer or from a URL. The next process is the face preparation process. In this process, the first step is to extract face from the image. This is done by marking the position of eyes and mouth. After the face is extracted, Multi-Resolution Active Shape Model is implemented to the image. After the facial feature points are extracted, user can adjust the facial feature points to the correct positions.

The caricature generation is a process of generating caricature from exaggerated facial feature points and base caricature. The process consists of selecting face characteristics (skin color, eye color, eyebrow color, lip color), shape exaggeration and image warping. The process of selecting face characteristic is used to build base caricature that is used in image warping process. Shape exaggeration is a process of exaggerating unique facial features. For example if the face has a big nose, the resulting caricature will have a bigger nose. In this process, the facial feature points is compared to the average of face model. If a face component is declared as normal, shape exaggeration process is not implemented to the corresponding face component. User can define the exaggeration rate for this process.

The last process is image warping process that warp the base caricature using the exaggerated facial feature points as destination points. The method that is used for image warping is triangular mesh warping.

Besides the main part, there are several programs to process the data that is necessary for facial feature extraction and shape exaggeration process. The processed data are the sample images and the facial feature points that is labeled manually to obtain the statistics that is necessary used for facial feature extraction and shape exaggeration process.

6. EXPERIMENTAL RESULTS

Portion of the testing in this paper, use the FERET database of facial images collected under the FERET program, sponsored by the DOD Counterdrug Technology Development Program Office [8,9].

One of the system web page can be seen in Figure 5 (gallery page).

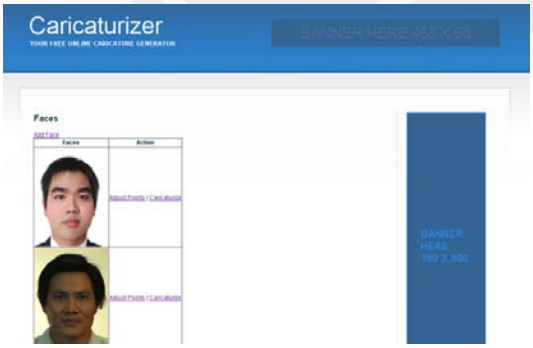


Figure 5. Website – Picture Gallery

In this page, user can select face from the gallery or upload a new image. After uploading a new image, user must mark the position of eyes and mouth then adjust the position of facial feature points a shown in Figure 6 and Figure 7.

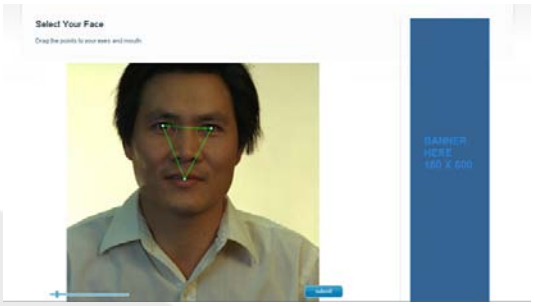


Figure 6. Marking The Position of Eyes and Mouth

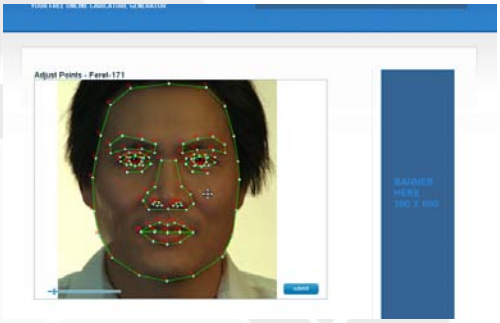


Figure 7. Facial Feature Adjustment

To generate a caricature, user must select a face in the gallery then select the corresponding face characteristics such as skin color, eye color, eyebrow color, lip color as shown in Figure 8.

Figure 8. Selecting Face Characteristics

Some resulting caricatures from the developed system with different exaggeration rate can be seen in Figure 9.

Original Face	No *Exaggeration*	*Exaggeration Rate* = 0.5	*Exaggeration Rate* = 1
			
			

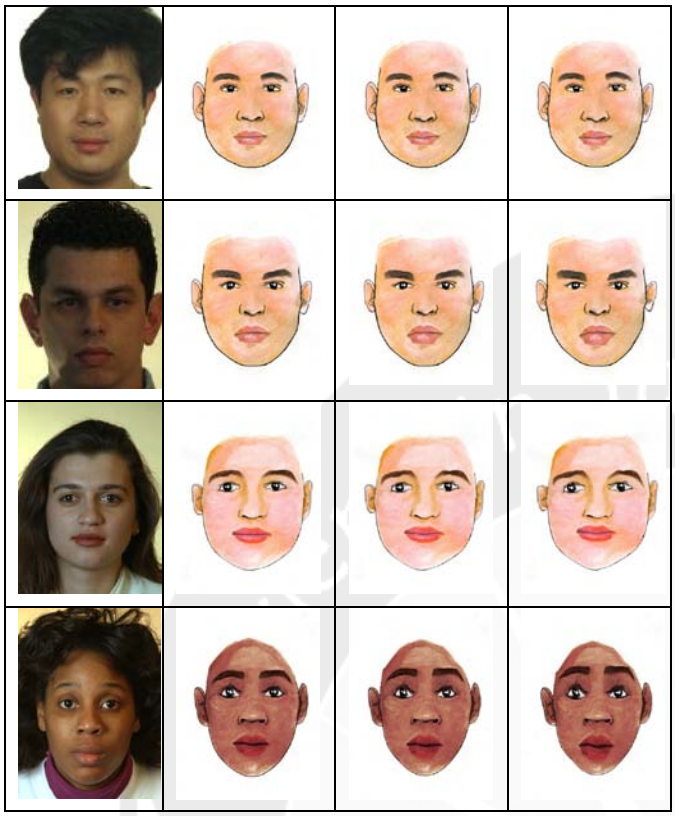


Figure 9. Experimental Results

From the resulting caricatures, it can be seen that the resulting caricatures has reflected the original faces. The exaggeration process has exaggerated the unique facial features. The higher the exaggeration rate, the more unique the facial features in the resulting caricature. However, if the exaggeration rate is to high, the resulting caricature will be distorted.

7. CONCLUSION

In this research, it has been developed an online caricature generator system to generate a caricature from a human face. The system can exaggerate the unique facial features. With higher exaggeration rate, the more unique the facial features in the resulting caricature. However, if the exaggeration rate is to high, the resulting caricature will be distorted.

8. REFERENCES

[1] Brennan, S. *Caricature generator*. Master's thesis, Cambridge, MIT. 1982.

[2] Chiang, P. Y., Liao, W. H., Li, T. Y. *Automatic caricature generation by analyzing facial features.* Proc. of 2004 Asia Conference on Computer Vision, Korea. 2004.

[3] Cootes, T. F. *An Introduction to Active Shape Models.* Image Processing and Analysis, Oxford University Press. 2000.

[4] Cootes, T. F., Taylor, C. J. & Lanitis, A. *Active Shape Models: evaluation of a multi-resolution method for improving image search*. Proc. of the British Machine Vision Conference. 1994.

[5] Gomes, Jonah, et al. *Warping and morphing of graphical objects*. The Morgan Kaufmann Series in Computer Graphics. 1999.

[6] Liang, L., Chen, H., Xu, Y. Q., Shum, H. Y. *Example-based caricature generation with exaggeration.* Proc. of 10th Pacific Conference on Computer Graphics and Applications. 2002

[7] Pereira, F. & Ebrahimi, T. *The MPEG-4 book*. Prentice Hall PTR Upper Saddle River, NJ, USA. 2002.

[8] Phillips, P. J., Moon, H., Rizvi, S. A., Rauss, P. J. *The FERET evaluation methodology for face recognition algorithms*, IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 22, pp. 1090-1104. 2000.

[9] Phillips, P. J., Wechsler, H., Huang, J. & Rauss, P. *The FERET database and evaluation procedure for face recognition algorithms*, Image and Vision Computing J, Vol. 16, No. 5, pp. 295-306. 1998.

Silny Approach to Edge Detection for Central Borneo Batik

Silvia
Informatics Engineering Atma Jaya University of Yogyakarta
Silvia310380@gmail.com

Suyoto
Informatics Engineering Atma Jaya University of Yogyakarta
Suyoto@mail.uajy.ac.id

ABSTRACT
Edge detection methods have been widely used for image processing / imaging. With this process the boundary between the background object can be determined properly. The number of edge detection methods that currently available can cause doubt in the decision-making appropriate which methods in accordance with the detected image conditions. Based on these issues, this paper aims to analyze the performance of edge detection by combining the method of Sobel and Canny, so getting a new method, the method can Silny as one alternative for image processing. Case studies will be conducted on Batik Bintik of Central Borneo.

Keywords
Edge Detection, Sobel, Canny, Batik Bintik of Central Borneo

1. INTRODUCTION
With the advance of information technology in the computer field, the more discoveries that obtained from the test results for further improve of the information technology. One of these information technology as we know the image processing, which has been applied to a number of areas, such as in medicine, biology, law, security and arts [5]. One of the main stages in the processing of images is the using of image edge detection process and for this there are several methods that have been widely used such as method of Robert, Sobel, Prewitt and Canny [6]. Of course, each method has its pros and cons each - respectively. If the edge-detection operator selection does not match, the result can lead to lack precision edge generated, other effects that occur can affect the process of further analysis.

The research has done base on these circumstances, by combining Canny and Sobell edge detection method, we can get a new methods that can be used as an alternative to edge detection image processing.

In this study the image to be used as a test is the motif of Batik Bintik from Central Borneo, this batik is one of the proud culture of Dayak etnic in Central Borneo.

2. EDGE DETECTION METHOD

2.1 Edge Detection

Edge is the change in the degree of gray intensity values of a sudden (large) within a short distance. The difference is that shows the intensity of the image details [5].

Edge can be oriented in one direction, and the direction are varies, depending on the intensity changes.

Edge detection is the first step to cover the information in the image. Edge characterizing object boundaries and therefore useful for the segmentation process and the identification of objects in the image. Edge detection operation goal is to improve the appearance of a boundary line or area objects in the image [6].

There are several techniques that can be used to detect the edge, they are [6] :

1. The first gradient operator, the example of the first gradient that can be used to detect the edge of the image are the centered deviation gradient operator, Sobel operator, Prewitt operator, Roberts operator, and Canny operator.
2. Second derivative operator, also called the *Laplace* operator. Laplace operator detect the location of the edge, especially on the steep edge of the image. At the edge of a steep, second's derivative have zero crossing, or at the point that have change of sign at the second derivative values, whereas on the sloping edge there is no zero crossing. An example is the *Laplacian* operator
3. Compass operator, is used to detect the edge of the various directions in the image. Compass operator is used to detect the edge of the display of 8 (eight) cardinal directions namely North, Northeast, East, Southeast, South, West, Southwest and Northwest. Edge detection is done by convoluted image with various mask compass, and then sought the value of edge strength (magnitude) of the largest and direction.

On the edge detection, we will apply the filter in the horizontal version of a bitmap, and for the other bitmap using a vertical version. And to combine them, use the formula pixels (1).

Formula pixels = sqrt (pixel1 * pixel2 * pixel1 + pixel2)…(1)

Edge detection works by finding the contrast of the image. This can be done several different ways, convolution filters do it by applying a negative weight on one side, and positive on the other side. This has the net effect of the trend towards zero if the values of the same, and the trend upwards as contrast exists. Excess of the generated images can be seen from the edge of a clear pedeteksian and the resulting noise reduced Filter will use the matrix, then detecting the edge of the image will be done to generate a new image better.

Many research has been developed to find the edge detection algorithm that produces better results, one of them is a new method which we develop based on edge detection algorithm of two existing and previously recognized. Silny method developed gradient method edge-detector combination of Sobel and Canny by using the identity matrix multiplication. In Silny method, the matrix X is used comes from the matrix x Canny method, while the matrix y derived from matrix y Sobel method.

Detection of the edge (Edge Detection) in an image is a process that produced the edges of image objects, the aim is [5] :

1. To mark the section of the detail image.

2. To improve the detail of the blurred image, which occurs because of error or the effects of the image acquisition process.

As it is known that the image acquisition results may contain a variety of characteristics [4], including noise and blur, then the solution of these problems is to develop other methods by adding the refining process (smoothing) before the edge detection process.

In broad outline the development of this method can be seen in Figure 1.

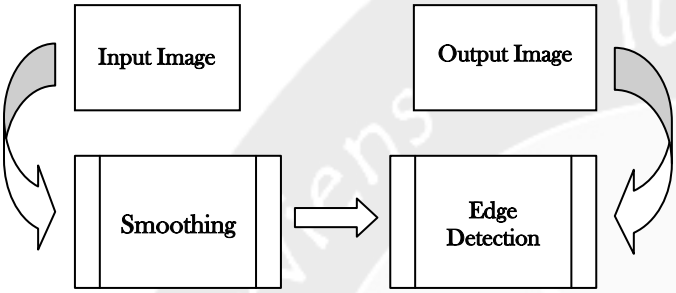


Figure 1. Chart Of Edge Detection Process Development

2.2 Sobel Method

Sobel method [2] have done the edge detection by giving attention for the vertical and horizontal edge.
Refining process used is a process of convolution of the window assigned to the detected image. In order to estimate the gradient of the middle window, the Sobel convolution using a 3x3 window, and the arrangement of pixels around the pixel (x, y) as the following chart (2).

p_1	p_2	p_3
p_8	(x,y)	p_4
p_7	p_6	p_5

The Sobel Convolution……(2)

So the value of the gradient is calculated using the equation below (3) and (4):

$sx = (p3 + cp4 + p5) - (p1 + cp8 + p7)$ ……(3)
$sy (p1 + cp2 + p3) (p7 + cp6 + p5)$ ……(4)

with the value of "c" is 2. It makes the Sobel operator matriks as follows (5) :

1	2	1
0	0	0
−1	−2	−1

(a)

1	0	−1
2	0	−2
1	0	−1

(b)

Sobel Matrix…….(5)

This method takes the principle of *Laplacian* and *gaussian* functions that are known as a function to generate FSH. The advantages of Sobel method is the ability to reduce noise prior to edge detection calculations.

2.3 Canny Method

Canny [1] proposed 3 criteria as base of the filters development to optimize the edge detection on the image bernois, namely:

a. *Good detection*,
This criterion aims to maximize the value of signal to noise ratio (SNR) so that all the edge can be detected perfectly or nothing was missing.
b. *Good localisation*,
Edge is detected on the actual position, or in other words that the distance between the edge position detected by the detector with the actual position is a minimum (ideally = 0).
c. *Low multiplicity of the response* or *"one response to single edge"* this detector give the actual edge.

Based on these 3 criteria optimization, Canny successfully produces the equation below :

$$h(x) = a_1 e^{\alpha x}\cos(\omega x) + a_2 e^{\alpha x}\sin(\omega x) + a_3 e^{-\alpha x}\cos(\omega x) + a_4 e^{-\alpha x}\sin(\omega x)$$

….(6)

but the equation is quite difficult to implement, so for the implementation Canny still using *Gaussian filter* to reduce noise and continued with the first derivative calculation and *hysteresis thresholding*.

Steps of Canny Edge Detection method:
1. Doing rarefaction (smoothing) to eliminate image noise by using Gaussian Filter.
2. Finding the gradient magnitude image to see the areas that have high spatial derivatives.
3. Determining the direction of the edge by using the inverse tangent of the gradient magnitude of Y (Gy) divided by the gradient magnitude of X (GX). Direction obtained from this calculation then mapped to 0, 45, 90, or 135 degrees according to the fourth degree of affinity with this direction.
4. Non-Maximum Suppression, removal of the values that are not the maximum.
5. Hysthresis do. Hysteresis using the two threshold T1 (lower threshold) and T2 (above threshold). If the magnitude is below T1, the point is set zero (become non-edge). If the magnitude is in the T2, so including the edge. If the

magnitude is between T1 and T2, the zero set unless there is a way (path) from point to point with a magnitude above T2
Matrix that used in Canny Method shows in (7).

-1	0	+1
-2	0	+2
-1	0	+1

Gx

+1	+2	+1
0	0	0
-1	-2	-1

Gy

Canny Matrix……..(7)

2.4 Silny Method

This new edge detection method is the combination between Canny and Sobel matrix method. In Silny method, the matrix X is taken from the value of the matrix X Canny, because this method can detect well-edge image that contain noise, while the matrix Y is taken from the value of the matrix Y Sobell, because this method is only good if used to detect images that do not contain noise . This research was conducted to see how well the results obtained in the event of a merger between these two methods.
Matrix that used in Silny Method shown in (8).

-1	0	+1
-2	0	+2
-1	0	+1

1	0	−1
2	0	−2
1	0	−1

Silny Matrix…….(8)

3. THE BATIK BINTIK OF CENTRAL BORNEO'S MOTIF

The Central Borneo's batik or often calls with Batik Benang Bintik have a very variatif motif and colors that can spoil the taste. A common motif is the Batang Garing (symbol of stems / tree of life for the Dayak community), Mandau
(traditional weapon of Dayak Etnics), Enggang/Tingang Bird (Borneo's Eagle), and Balanga (The Dayak's Jar) or the combination of them.
Batik cloth used not only for formal events, but also can be used on the non-formal events. In this test used Batik Bintik with Batang Garing motif, because this motif is more frequently used by the Dayak tribe in Kalimantan [8].
Examples of batik spots can be seen in Figure 2.

Figure 2. Example of Batik Bintik Motif

4. RESEARCH METHODS

Method of research is done by entering one by one matrix Sobel operator, Canny, and Silny which then continued to include the image in the form of Batik Kal-Teng Spot Batang Garing motives into software. The results of these three methods and then compared.
The steps is shown by the Figure 3.

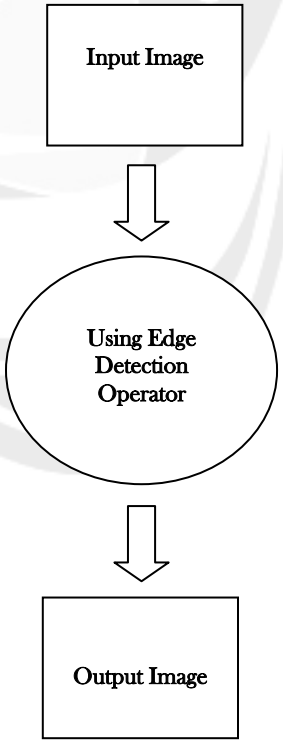


Figure 3. Steps of Image Edge Detection Reseach Chart

5. EXPERIMENT RESULT

Pictures below is the image of the edge detection results using the method of Sobel, Canny method, and the combined method of Sobel and Canny methods (Silny) :

Batik motif that will be tested as shown in Figure 4 .

Figure 4. Bintik Motif (Batang Garing Motif)

5.1 Testing with Sobel Method

Matrix that will be entered in to the software is matrix of the Sobbel Operator, with 3 x 3 windows. Image resulted by Sobel Method can be seen in Figure 5.

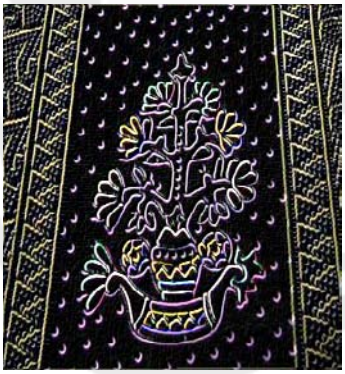


Figure 5. Image resulted by Sobel Method

Analysis :

- Motif spots / dots only appear on the right side (not a complete form),
- Edge detection for the main image appear on 1 side
- The shape curve on the edge of the image can not be unified whole.
- Triangle form, the right edge detection is not visible and not straight.
- Picture of straight line can be seen clearly

5.2 Testing with Canny Method

Matrix that will be entered in to the software is matrix of the Canny Operator, with 3 x 3 windows. The result of Canny method implementation on Batik Image can be seen in Figure 6.

Figure 6. Image resulted by Canny Method

Analysis

- Detection of left and right edge is visible and thinner, because every edge / line, represented by only one response / pixel.
- Motif spots / dots only appear on the left side (not the shape intact).
- Edge detection for the main image appear on 1 side
- Triangle form, edge detection for the left and right are not visible and not straight

5.3 Testing with Silny Method

Silny matrix is the combination of matrix Sobell and Canny.

In Silny method, value of X taken from the Canny method because this method has been shown to have a noise level lower than the other methods. While the matrix Y is taken from the Sobell methods.

Matrix that used inSilny Method shows in (9).

-1	0	+1
-2	0	+2
-1	0	+1

1	0	−1
2	0	−2
1	0	−1

Silny Matrix…….(9)

The result of Silny method implementation on Batik Image can be seen in Figure. 7.

Figure 7. Image from the Silny Method

Analysis :

- Because Silny Method is a combination from two methods, namely Sobel and Canny method, then the edge detection method produces canny generated imagery is also a combination of both methods. Edge detection on the main picture (center) looks much thicker and form a more perfect arches (intact).
- Spots / dots image that previously only appeared on the right or left, with this method results are more complete (full round).
- Straight line image can be detected better.
- However, this method is not perfect either, because it still looks a the noise is still apperared, looked at the arch shape on the bottom rod Garing motive is not clear / not detected properly as well as the triangle pattern on the edge of the imaging area.

Edge detection results using the method of Sobel, Canny and Silny are not provide optimal results, it is proved by the noise that contained in the image is still well detected. In addition, many of the images are not detected, this analysis resulted in the detection of the edge image to be inaccurate.

On the edge detection results using the Silny method seems that there is improvement from the image, though not very optimal. Some images that are not whole, because of this combined method, can be displayed properly

6. CONCLUSION AND RECOMMENDATION

6.1 Conclusion

- Based on the results of experiments by entering matrix Sobel, Canny, and Silny into the software, can be seen that each method has its pros and cons.
- In the Sobel method, edge detection imaging is more focused on the right, while the Canny method is more focused on the left. As a result, imaging with Sobel method would be better on the right, the opposite applies for the Canny method.
- Because Silny Method is a combination of Sobel and Cany Method, seems like there are improvement of imaging, though not perfect yet.

6.2 Recommendation

Because Silny is a combination methods, the noise generated by each method also incorporated, and slightly curved lines still not detected. Therefore further research is still needed so the process of combining these two methods can be minimized noise values and obtained better results in imaging and more accurate.

7. REFERENCES

[1] Canny, J.1986. *A Computational Approach To Edge Detection.* IEEE on PAMI. vol. 8, pp. 679-697

[2] Sobel, I., 1990. *An isotropic image gradient operator* . In H. Freeman, editor, Machine Vision for Three-Dimensional Scenes, pages 376--379. Academic Press.

[3] Sobel, I., 1979. *Camera Models and Perception,* Ph.D. thesis, Stanford University, Stanford, CA.

[4] Madenda, S., R. Missaoui, J. Vaillancourt & M. Paindavoine. 2006. *An Optimal Edge Detector for Automatic Shape Extraction*. SITIS

[5] Febriani, Lusia. 2008. Analisis Penelusuran Tepi Citra Menggunakan Detektor Tepi Sobel dan Canny. Fakultas Ilmu Komputer dan Teknologi Informasi, Universitas Gunadarma.

[6] Munir, Rinaldi. Pengolahan Citra Digital. Informatika. Bandung. 2004.

[7] Keren, D., Osadchy, M., & Gotsman. C. (2001). Antifaces: A novel, fast method for image detection. *IEEE*

[8] Abidin Y.Kumpulan Motif Batik Printing Khas Dayak Kalteng. SMK Negeri 4. Palangkaraya. 1997

[9] Milan, S., Vaclav, H., & Roger, B. (2002). *Image processing analysis and machine vision*. London: Chapman and Hall, 255-280.

[10] Gonzalez, R., & Woods, R. (2002). *Digital image processing* (2nd ed.). Prentice-Hall Inc. 567-612.

[11] Chang-Huang, C. (2002). *Edge detection based on class ratio*. 152, sec.3, Peishen Rd., Shenkeng, Taipei, Taiwan, R.O.C.

[12] WilliamK.Pratt,"DigitalImageProcessing_WilliamK.Pratt3rd" ,chapter15.

Cattle's Cost of Goods Sold System Information At CV Agriranch

Lily Puspa Dewi
Informatics Dept. Petra Christian University
Siwalankerto 121-131
Surabaya, Indonesia
62-61-8439040
lily@petra.ac.id

Yulia
Informatics Dept. Petra Christian University
Siwalankerto 121-131
Surabaya, Indonesia
62-61-8439040
yulia@petra.ac.id

Anita Nathania
Informatics Dept. Petra Christian University
Siwalankerto 121-131
Surabaya, Indonesia
62-61-8439040

Doddy Hartanto
Informatics Dept. Petra Christian University
Siwalankerto 121-131
Surabaya, Indonesia
62-61-8439040

ABSTRACT

CV. Agriranch is currently using a manual system and simple database program to manage the execution of purchasing, sales, and production. The system, however, is ineffective to determine the loss and profit gained by the company per period.
Therefore, a new application was designed to assist with the workflow in CV. Agriranch. The application supports the lifecycle in the company through purchasing, sales, cost of goods sold calculation, and monthly report of inventory status. The application is using PHP as programming language, and Microsoft SQL Server 2005 as data storage media.

Based on the experiment, 80% of the users stated that the features on the application are sufficient to fulfill the needs of the company.

Keywords

FIFO, Cost of Production, Cost of Goods Sold, Cattle.

1. INTRODUCTION

CV Agriranch, located at Raya Driyorejo km. 19,3, Gresik, Jawa Timur was founded in 2007 by Mr Kim Chance MLC, West Australia Minister for Agriculture and Food. This company builds farm and cattle feedlot. Agriranch has a commitment to procure high quality beef cattle both from domestic resources and Australia. Cattle are raised intently and fed with high nutrition feed for periods ranging from 90 to 100 days to achieve specific market requirements.
Over the years, purchasing, sales and data recording are conducted manually using Microsoft Excel. These facts posed a challenge made it difficult for the producer to calculate "cost of goods sold" for the cattle. Besides, the manual system causes difficulties in generating detailed reports at the end of period. To resolve these problems, a new computerized system is needed to help Agriranch provide accurate and detailed data and report at the end of period. In the end, this new system can raise more profit for Agriranch.

2. COST OF GOODS SOLD

In a production of the goods, there are two types of costs, i.e. costs of production and non-production costs. Production costs are costs incurred for processing raw materials into finished products., while the non-production costs are the costs incurred for non-production activities, such as marketing activities and administration [1].

Cost of goods sold can be categorized into three types. First, direct materials are all materials that form an integral part of the finished product and that can be included directly in calculating the cost of the product. Examples of direct materials are the lumber to make furniture and the crude oil to make gasoline [3]. Second, direct labor is labor that converts direct materials into the finished product and the cost can be applied accordingly to a specific product [3]. Third, Factory overhead-also called manufacturing overhead which includes all manufacturing costs except those accounted for as direct cost, i.e., direct material and direct labor [3]. Below is the formula to assign cost of goods sold:

Cost of goods sold =
Beginning Work In Process inventory +
manufacturing cost – Ending Work In Progress inventory

3. SYSTEM ANALYSIS

In general, the company business process starts from cattle purchasing. At the farm, incoming cattle will be weighed and checked for health. Then, each cattle will be given ear tag as its identity. The cattle will be weighed monthly and any weight changes will be recorded. It is useful to know whether the weight gain of cattle balances with the amount of food. For selling process, buyer directly comes to the farm and choose the cattle they want to buy. When both parties have reached an agreement in regards to the price, cattle health, payment methods, etc., the buyer can take the cattle directly. The cattle feedlot production cost includes the employee cost, electricity, water, feed, vitamin,injection and medicine. At Agriranch, the actual profit of this farm was not known exactly because there was no proper guideline and system to a record all of the production costs. Reports are available only to the extent of price and number of cattle sold each month and the weight gain of the cattle. Document Flow of cattle purchasing can be shown as figure 1 below.

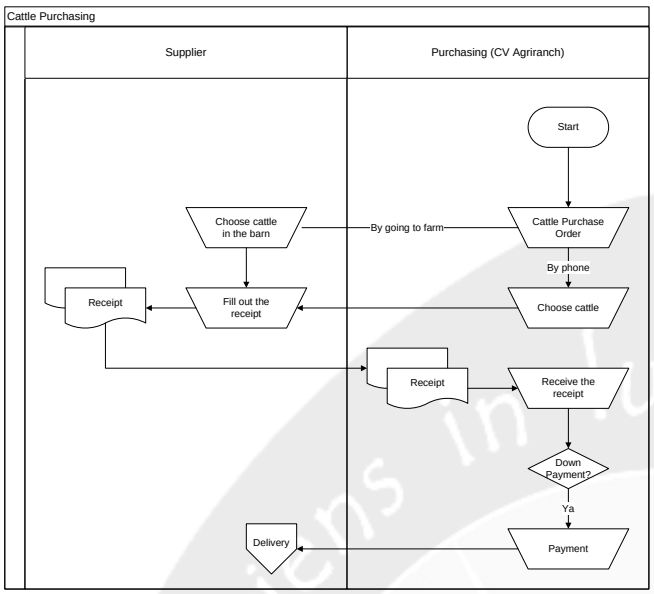


Figure 1. Document Flow – Cattle Purchasing

4. DATA FLOW DIAGRAM

Context Diagram Design can be seen on figure 2 below. There are 4 external entities that deliver input and output to the system, i.e. a. Supplier who receives purchasing order from CV Agriranch. Any materials are inputed to the system. B. Customer who buys cattle from the farming. The sold cattles are inputed to the system. C. Manager who obatins reports from the system. D. Staff who is responsible to breed and sell the cattle.

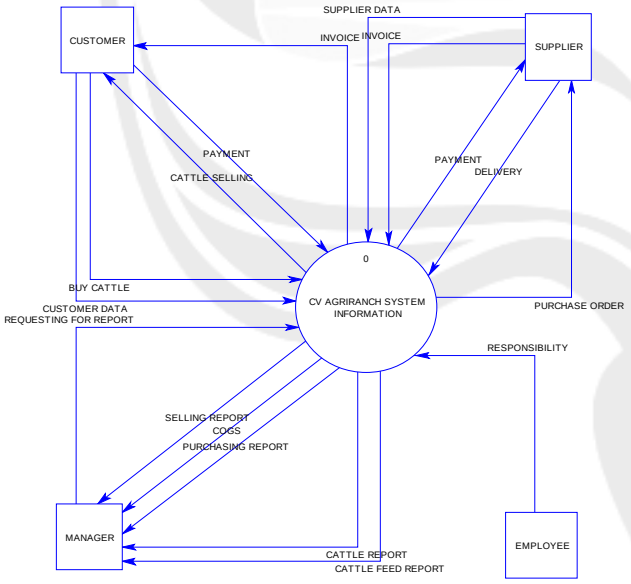


Figure 2. Cattle's Cost of Goods Sold Context Diagram System

This application consists of four processes, i.e. :

1. Purchasing. In this process, purchase order is generated and the purchase transaction between the buyer and supplier occurs.
2. Sales. This process covers cattle sale to the customer.
3. Breeding. This process consists of 4 sub processes, i.e. cattle weighing, cattle care and handling such a administering medicine, vitamin and injection, cattle feeding and calculating breeding cost . Breeding process will give some output like cattle weight, cattle feeding consumption, and medicine consumption. These output will affect to the cost goods sold calculating. As for the calculation of breeding costs are as follows:

$$\text{Labor cost} = \frac{\text{Labor cost in 1 month x month}}{\text{Numbers of days x cattles}}$$

$$\text{Electricity cost} = \frac{\text{Electricity cost in 1 month x month}}{\text{Numbers of days x cattles}}$$

$$\text{Water cost} = \frac{\text{Water cost in 1 month x month}}{\text{Numbers of days x cattles}}$$

$$\text{Overhead cost} = \frac{\text{Overhead cost in 1 month x month}}{\text{Numbers of days x cattles}}$$

4. Reporting. This process will provide purchasing, sales, stock and cost of goods sold report.

5. IMPLEMENTATION

The evaluation starts from purchase order, cattle receiving, medicine supply purchasing, cattle breeding, sales and cost of goods sold calculation.

5.1 Cattle Purchase Order

Figure 3 below show the interface to make a purchase order. For example, an order to PT SAPISEHAT of three types of "Brahman" cattle with an average weight of 190 kg and the average price of Rp. 15.000,- and two types of "Brahman Cross" cattle weighing on average 180 kg and the average price of Rp. 12.000,-.

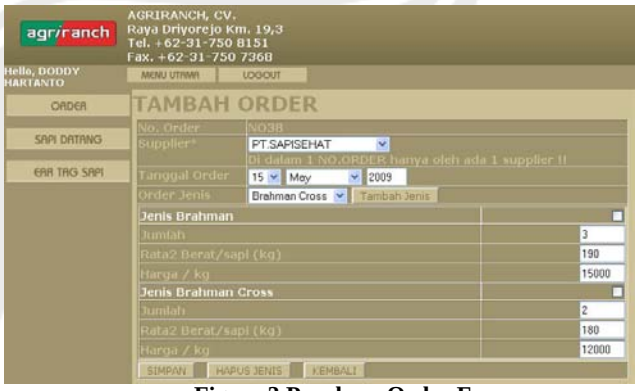


Figure 3 Purchase Order Form

5.2 Cattle Receive Form

When the cattles arrive at the farm, the cattleman will record the data. Figure 4 show 3 suppliers sent three types of Brahman with the total price of Rp. 9,000,000,- and an additional fee of Rp.150.000,-.

Figure 4 Receive Form

5.3 Medicine Supply Purchasing

For medicine purchasing, the transaction will be stored in this form below (figure 5). For example, the farm wants to buy medicine at PT OBAT KIMIA, a bottle of Sangobion that costs Rp 15.000,-. per each. This form can be seen as follow.

Figure 5 Medicine Supply Purchasing Form

The stock card of Sangobon will be added to one bottle or 50ml (1 bottle equals to 50ml). One bottle of Sangobion costs Rp 15.000, therefore 1 ml would cost Rp15.000/50 = Rp 300,-. The form can be seen as follow.

KARTU STOK OBAT SANGOBION

NO. STOK	TANGGAL	JUM_MASUK	JUM_KELUAR	SISA OBAT	SISA OBAT FIFO	HARGA MASUK	HARGA KELUAR
NSO32	12 Apr 2009	250	0	250	0	300	0
NSO34	12 Apr 2009	0	100	150		0	300
NSO35	12 May 2009	100	0	250	80	200	0
NSO36	12 May 2009	0	70	180		0	300
NSO37	13 May 2009	0	100	80		0	280
NSO38	15 May 2009	50	0	130	50	300	0

Figure 6. Medicine Stock Report

5.4 Cattle Care and Handling

Medicine can be administered when cattle are sick. Example, May 16, 2009, a cow with identification # NS95 needed 50ml of Sangobion. It can recorded as follows.

Figure 7. Medicine usage

Card stock of Sangobion will be reduced by 50ml with the price of Rp 200,-/ml. This price was obtained from medicine usage NS035 (sangobion). The stock report of sangobion can be seen in figure 8

KARTU STOK OBAT SANGOBION

NO. STOK	TANGGAL	JUM_MASUK	JUM_KELUAR	SISA OBAT	SISA OBAT FIFO	HARGA MASUK	HARGA KELUAR
NSO32	12 Apr 2009	250	0	250	0	300	0
NSO34	12 Apr 2009	0	100	150		0	300
NSO35	12 May 2009	100	0	250	30	200	0
NSO36	12 May 2009	0	70	180		0	300
NSO37	13 May 2009	0	100	80		0	280
NSO39	16 May 2009	0	50	30		0	200

Figure 8. Medicine Stock Report

5.5 Cattle Selling

For example, user sells a cow -to Edo with identification number 004. That cow weighs 191 kg and the price per kg is Rp. 18 000, The process can be seen in figure 9.

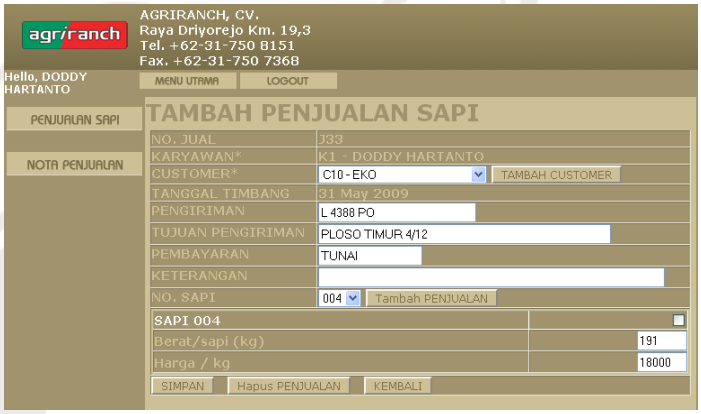


Figure 9 Cattle Selling

5.6 Reporting

This is Cost of Goods Sold report in April 2009. All sales in April 2009 will be calculated with conversion cost in the same month.

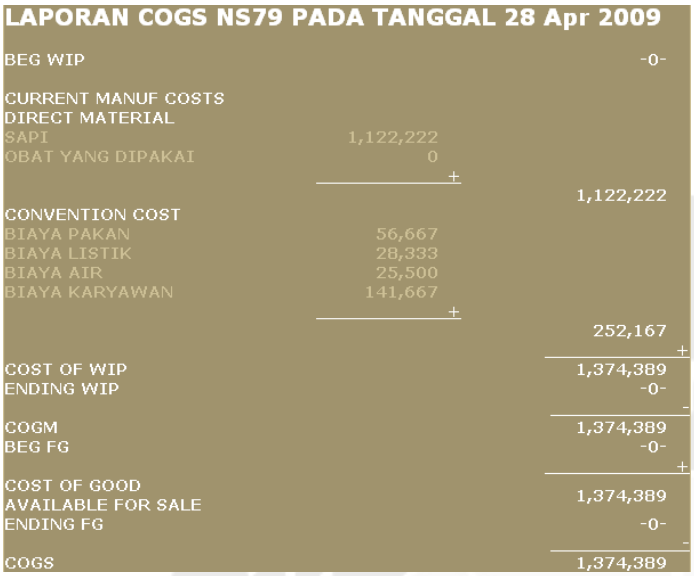


Figure 10 Cost of Goods Sold Report

6. EVALUATION

The evaluation of the application of this program is done by analyzing the questionnaires of the five users who carried out tests on this application. This evaluation is done through a user rating given to the criteria mentioned in table 1.

Table 1. Questionnaire Result

Result / Question	1 (poor)		2 (fair)		3 (good)		4 (very good)		5 (excellent)	
User Friendly	0	0%	0	0%	0	0%	1	20%	4	80%
System Suitability	0	0%	0	0%	0	0%	1	20%	4	80%
Problem Solution	0	0%	0	0%	0	0%	4	80%	1	20%
Process' Accuracy	0	0%	0	0%	1	20%	1	20%	3	60%
Application Interface	0	0%	0	0%	0	0%	4	80%	1	20%
Benefits Application	0	0%	0	0%	0	0%	1	20%	4	80%

7. CONCLUSION

This Cost of Goods Sold application runs well and helps CV Agriranch improve company performance gained on each month by providing actual expenses and cattle detail record.

Based on the evaluation, this application has fulfilled the user's needs. This is shown by the fact that 80% of respondents provide positive feedbacks when answering question number three on suitability of the features created to the needs of users. In addition, this application can be proven as beneficial for the company by looking at the total percentage of answers to question number six, which asked about how large the benefits of this application for a company that is 80% of users replied very useful

8. REFERENCES

[1] Carter, W.K. & Uzry, M.F. (2004). *Akuntansi Biaya* (13th Ed). Jakarta : Penerbit Salemba Empat.

[2] Mulyadi. (2005). *Akuntansi Biaya* (5th Ed). Yogyakarta : Unit Penerbit dan Percetakan Akademi Manajemen Perusahaan YKPN.

Compensation Method for Internet Grids Using One-to-Many Bargaining

Andreas Kurniawan
New Frontier Solutions Pte. Ltd.
Jl Jend. Sudirman Kav 45-46 Floor 24
Jakarta 12930, Indonesia
+62-21-57951288
admin@andreaskurniawan.web.id

Pujianto Yugopuspito
Universitas Pelita Harapan
Jl M.H. Thamrin Boulevard
Tangerang 15811, Indonesia
+62-21-5460901
yugopuspito@uph.edu

Johan Muliadi Kerta
Universitas Bina Nusantara
Jl Kebon Jeruk Raya no. 27
Jakarta 11530, Indonesia
+62-21-53696969
johanmk@binus.edu

ABSTRACT

Well-known grid computing project, Folding@home, is recorded in Guinness World Records because using 4.223.713 computers and producing 5 petaflops by implementing volunteer computing. Total internet users in the world are 1.596.270.108 computers, and it means that Folding@home used only 0.265% from all of the computers available. The solution to increase the total resource used is using compensation method. One-to-many bargaining proven to save as much as 24.17% compared to the task provider expenditure using a one-to-one bargaining. From the cost side, the proposed model called CCCM which takes bandwidth into account. The research is successfully implemented in Rocks Clusters 5.1 and Tachyon 0.98.7 using MPICH.

Keywords

grid computing, compensation method, one-to-many bargaining, bandwidth.

1. INTRODUCTION

Grid computing is a combination of several computers to solve a problem at the same time, where the problem is usually a matter of technical knowledge and which require computer processing cycles or large amounts of data access [1]. One of grid computing strategy is to use software to share some part of the program to several computers, which usually amounted to thousands. Grid computing is better than conventional supercomputers, because every component of the grid can be purchased separately and then combined using a computer network that can produce similar computing resources with a multiprocessor supercomputer, but with cheaper prices.

One well-known grid computing project is Folding@home. The project was recorded in the Guinness World Records as the biggest distributed computing cluster [2]. This project uses 4,223,713 computers and produces 5 petaflops [3]. In order to achieve that, Folding@home applies volunteer computing. Any person may voluntarily donate the computer's performance to help this project so the computation process can be done more quickly than before.

The need for grid computing has increased from time to time. In computer graphics, there is a technique called ray tracing that is used to produce images by tracking the path of light through pixels in an image. This technique requires high computing. Film and television industry is now starting to use ray tracing to describe the picture slowly [4].

By knowing this situation, grid computing projects that are previously carried out by organizations non-profit started to move toward commercial. This is supported by the fact that the cost to build a grid computing infrastructure is not cheap. A grid infrastructure costs approximately $225,000 for 32 clusters with its system administrators, resources, and cooling for a period of 5 years [5]. There are 2 companies that already commercialize grid computing. They are Amazon's EC2 with prices $0.10 per instance-hour, and the Sun Grid with price $1 per CPU-hour [6].

Grid computing projects always require large resources, even the infrastructure that has been built by large companies were not able to serve customer demand. Amazon EC2 ever experienced API outage on February 15, 2008 [7]. It is almost impossible to propose a grid computing infrastructure that can serve all customer needs, because it will need a lot of money.

Commercial companies cannot use volunteer computing like Folding@home, because the main purpose of the company is for profit, so there would be a few people who are willing to help. Besides volunteer computing, companies can implement compensation system for what they have contributed. If this compensation system to function properly, it is not impossible that this system will be able to beat the volunteer computing. That is possible because the number of Internet users in the world is 1,596,270,108 computers [8]. Folding@home is now just using 0.265% of the total number of the world's computers that are connected to the Internet. With the compensation, then the people will be more willing to provide resources, because they think that they are not loss for what they have, but they will make a profit from the compensation.

Compensation method is usually used in internet marketing. Until now there are 4 methods of compensation: cost per click, cost per action, cost per sale, and cost per mille [9]. Four of them are not suitable when applied to grid computing. A new method should be proposed uniquely, because grid computing is not a form of human interaction, but more to the computer resources.

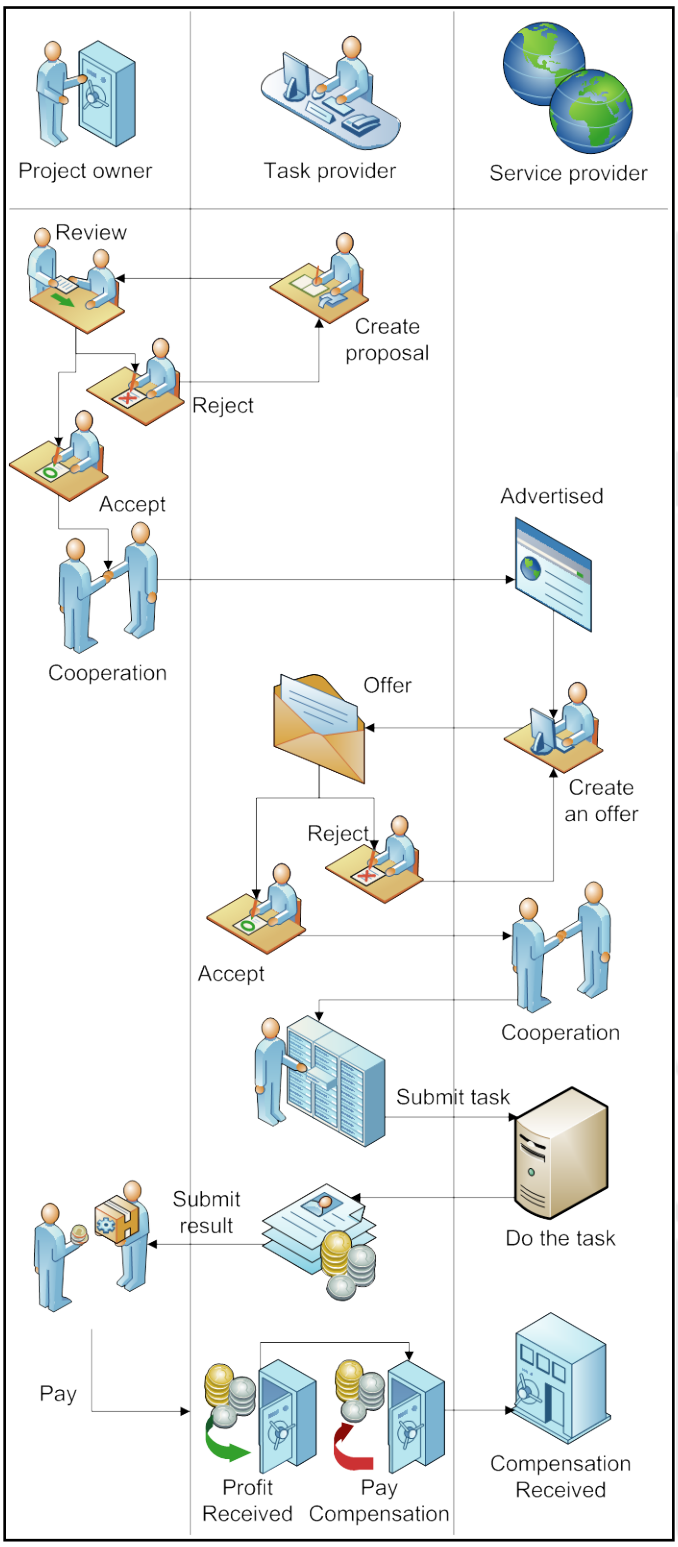


Figure 1. Cooperation Flow for Project Owner, Task Provider, and Service Provider

2. COOPERATION FLOW

There are three players involved in the negotiation: project owner, task provider, and service provider. Figure 1 shows the cooperation flow from the beginning of the project until the compensation is paid.

Initially, the project owner has a project that requires high computing resources. Task provider will create a proposal and propose it to the project owner. Project owner will review the entire proposal and select a task provider. If project owner refuses the task, then the task provider can revise the project proposal that have been made before if they still want to deal with the project owner.

After the project owner cooperates with the task provider, the next task for task provider is to find resources to work on projects that have been received. Service providers that are interested to the project will make an offer. Bargaining process will occur between task provider and service provider. The process of bargaining is interesting, and it will be discussed more detail later. Bargaining is not just occured between two players, because the number of service providers could be more than one. This process is called a one-to-many bargaining.

If there is a deal between the task provider and service provider, then the next task for task provider is to send tasks to service providers. Service providers will do the job with its resources. Processing results will be sent to the task provider. Task provider will collect the results from each service providers and combine it to be sent to the project owner.

Project owner who receives the results of computation will provide rewards to task providers who work as promised as written on the contract. Task provider receives profits from task execution results, and will provide compensation to service providers. Service providers will receive compensation as promised in the beginning, and the process has reached the end.

3. MODEL EVALUATION

In this research, there are two models that need to be discussed: Pricing Model and Cost Model.

Pricing model is used to determine which decisions will be taken by task provider and service provider during the negotiation. Discussion of the pricing model is needed to meet the needs of service provider individual preferences. Service providers can come from anywhere using the Internet.

Cost model is used to determine the amount of compensation that task provider should pay for the resources. The results of the negotiations based on the price model will be considered in the formulation of cost models.

Table 1 shows the comparison of some previous research on implementation of game theoretic in grid computing.

Table 1. Scheduling system for grid computing research

Grid Scheduling Systems	Game Theoretic Model	Discuss about
Spawn [10] and Popcorn [11]	Auction Model	Sealed-bid, second-price auctions.
Nimrod-G [12]	Bargaining Model, Posted Price Model	Deadline and budget constraint.
Mungi [13], MOSIX [14] and Nimrod-G	Commodity Market Model	Rent.
Rexec and Anemone [15]	Bid based Proportional	Utility function.

	Resource Sharing	
SETI@Home, Condor [16] and MojoNation [17]	Community, Coalition, Bartering	Content sharing.
Mariposa [18]	Tender / Contract-Net Model	Budget-based processing.
Ghosh [19]	Non-cooperative Bargaining Model	Pricing strategy.
"This paper"	One-to-many Bargaining Model	Compensation method.

Table 2 shows the comparison of some previous research on cost model in grid computing.

Table 2. Cost model research

Grid Scheduling Systems	Cost Model	Cost Factor	Used For
Grosu [19]	COOP	*CPU cycle*	*Distributed System*
Ghosh [20]	OPTIMAL	*CPU cycle* dan *price per unit*	*Mobile Grids*
"This paper"	CCCM (Computational and communication cost model)	*CPU cycle*, *bandwidth*, dan *price per unit*	*Internet Grids*

4. ARCHITECTURE

Price model used in this paper is using game theoretic framework, especially bargaining model. The Players who are involved in this case are task provider and service providers. The amount of task provider is one, while the amount of service provider can be more than one. Service providers come from the internet to compete contributing in this project in order to get compensation. The concept of incomplete information between the players is to make sure that each player does not know of another assessment, which are the highest price from the task provider and the minimum price desired by the service provider. The illustration about the interaction between task provider and several service providers is illustrated in Figure 2.

In a grid environment, job providers will try to get some resources that is offered by service providers that appear on the internet. Task provider have n round of negotiation to get the price per resource from Q resources. At time t, total available service providers is denoted as n(t), so that n (t) = Q. Figure 2 and figure 3 show that there is one task provider and several service providers. If we can model the relationship between task provider and service providers as a bargaining game BG_j, for $1 \leq j \leq Q$, and manage their relationships properly to produce the game output, then the scenario can be considered as one-to-many bargaining with delay d as the time to collect offers from service providers.

Results from game after delay d denoted as $\Gamma(q)$ where q is the possibility of negotiations will end at a certain time. Bargaining process uses the same concept as alternating offer. A pair (λ, γ) is defined as the strategy used to produce (x^t, t). If x is defined as a result of the first offer, and y is defined as a result of the

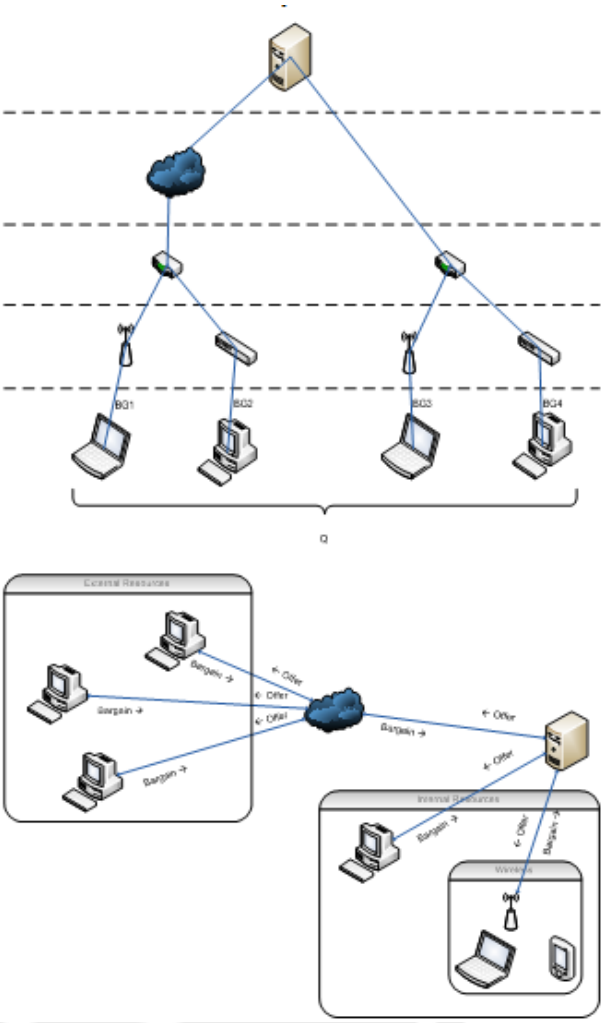


Figure. 3 Game Notation

second offer, then t is the time associated with these results. (λ, γ) will lead to an agreement (x^t, t) with the probability $(1-q)^t$ and probability to fail $1-(1-q)^t$.

The term bargaining according to Nash [21] assumed the condition as follows:

a. The conflict of interests of both players.
b. Both players have the power to end the bidding.
c. Each offer cannot be terminated by only one player.

Figure 4 shows an illustration of bargaining protocol between task provider and service providers.

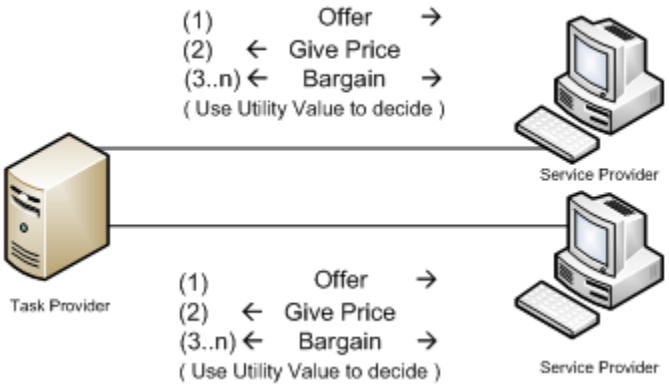


Figure. 4 Bargaining Protocol

5. ONE-TO-MANY BARGAINING

Bargaining process starts from task provider making an offer to all service providers. Service providers can accept, bargain, or reject it. If the offer is accepted by the service provider, then the bargaining is ended and there is a deal. If the offer is rejected, then the bargaining process is complete and there is no agreement between the two players. If the offer is bargained, then the service provider must submit a new proposal to task provider. Task provider will collect proposals from service providers and replies it within a certain time frame. From the results obtained, the task provider can determine which proposal is accepted, rejected, or bargained. Bargaining process will continue until one player accept or reject the offer.

Ghosh offers a bargaining solution which describes the space that can be accepted by each player based on the time and resources. The bargaining solution space presented by Ghosh can be seen at Figure 5.

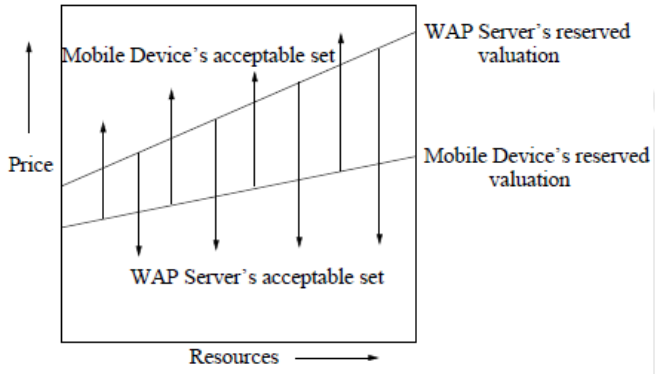


Figure. 5 Bargaining Solution Space from Ghosh

Ghosh's bargaining solution, which is using one-to-one bargaining, needs some modifications to be used on one-to-many bargaining. Bargaining solution space proposed by this paper has considered the threshold for maintaining cooperation between service providers. Function owned by the task provider have also considered the project deadline. A picture of reservation value for each player based on two variables, price and time is illustrated at figure 6.

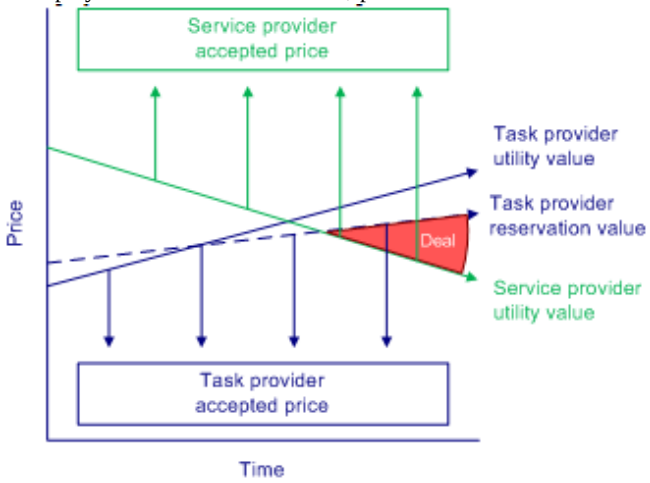


Figure. 6 One-to-many bargaining solution space

Utility value is a value to determine wheter the player should accept or decline the offer. The utility value for service provider is the final value after divided by CPU cycles to fulfil the fairness. The value can be formulated as follows:

$$\delta_i = \frac{p_i}{a_i}$$

where δ_i is the price per CPU cycle, p_i is the price preference, and a_i is the CPU cycles of service provider i.

With the value of per CPU cycle, then the value held by each service provider can be compared. From Figure 6, the service providers price tend to be down from time to time, because with the pressure of time, service providers are assumed to use Time-dependent threshold strategies [22]. In this case the task provider use Time-dependent and responsive threshold strategies that have two lines, the utility value and the reservation value. For task provider, utility value is the lowest value offered by service providers, while the reservation value is the maximum value from task provider to avoid any coalition between the service providers. Task providers also have the time pressure to complete the project on time, so reservation values will increase over time. Task provider will try to complete the task on time, because the patience of service providers will be tested, so the task provider will get more profit.

6. PRICE MODEL

Based on the assumption of strict limitations-game approach theoretic [23], the game can be characterized into the following rules:

a. Each player is rational, which each have a preference level and always choose the best. Ghosh proposed formula for calculating the expected utility as below:

   Expected utility = E [Surplus] = (reserved valuation of w - standard price) x probability (standard price)

This paper proposed a new formula for task provider by adding considerations of project time and price offered by each service provider. Changing of ghosh formula is to adjust to the process of negotiation using one-to-many bargaining. From the provider side of the task,

$$R = x + \left(1 - \left(\frac{f}{h}\right)\right)(y - x)$$

where R is the reserved valuation, x is the minimum price, y is the maximum price, f is the number of days before the project ended, and h is the total project days.

Determination of threshold value by using the minimum price and maximum price is based on the price agreed between the project owner and the task provider. The maximum price that is included in this model should be below the price of commercial grid, Amazon EC2 and Sun Grid. Amazon EC2 sell 1 GHz for $0.10 per hour, while the Sun Grid sell 3 GHz for $1 per hour. In this case Amazon EC2 3 times cheaper than sun grid. So in this case we will compare it to the Amazon EC2. If we change the price from Amazon EC2 CPU becomes per cycle, then the price is $0.0001. For the calculation of compensation per hour, the maximum value of the above formula should be $0.0001 or below, in order to compete with Amazon EC2.

$$V = \min(R, \min(w))$$

where V is the value of minimum valuation, and w is the offer of the service provider. By knowing equation (1) and (2), expected utility value formula can be written as below:

$$E = 1 - \frac{(m - V)}{(z - V)}$$

where E is the expected utility value, m is the offer, and z is the highest offer. From the service provider side,

$$E = (m - R)s$$

where E is the expected utility value, m is the offer, and s is the possibility that the proposed price will be accepted.

b. If the offer is rejected, then both the service provider and task provider will reduce their preference base on time that increase the possibility that the bid will be accepted by the other player. Task provider will reduce the reserved valuation based on the project deadline after reaching a certain period, while service providers will reduce the probability exponentially, by knowing the time pressures on service providers.

c. Both players do not remember the previous experience, so in this case there is no learning process occurs. Task providers will only see the value based on time deadlines and minimum offer. On the other hand, the service provider only knows how long the bargaining process has occurred.

7. COST MODEL

Previous research conducted by Ghosh consider the CPU cycles as a factor for the cost calculations. The price negotiated also included as an agreed price. Below is the model proposed by Ghosh:

$$C = \sum_{i=1}^{n} \frac{\Omega_i p_i \theta_i}{\Phi(\mu_i - \theta_i)}$$

This paper proposed a new model by adding bandwidth as a communication cost. In the previous architecture, internet connection is not needed. In internet grids, internet connection should be considered. Each task and its result will be sent over the internet, so the service provider should provide the connection. As a task provider, it is needed to calculate the communication cost in the model.

In a system there is n service providers. Task provider already know the price for service provider i is j_i with processing speed μ_i and delivery s_i. Service providers have agreed that there are two resources used, the CPU cycles and bandwidth.

Compensation for CPU cycle will be calculated based on time usage, while for the bandwidth is calculated based on the total data size. This type of compensation is frequency usage. CPU cycles are being utilized by the task provider, so that calculations based on time is more appropriate than by total usage. Bandwidth used for delivering computation results are not always used. Results are sent after the computation process is complete. Therefore, the task provider, the calculation of compensation cost based on usage will be more efficient than time-based.

Generally, compensation is calculated per unit hour, while the negotiations conducted with a range longer than compensation. Equation (1) which is used to calculate the cost per CPU cycle, $\delta_i = \frac{p_i}{a_i}$, need to be modified in order to model the time unit costs by more detail to:

$$o_{ij} = \frac{l_i}{a_i t}$$

where o_{ij} is part of the number of CPU cycles, and l_i is the number of CPU cycles generated by the service provider i during the time t at part of time j.

$$\alpha_{ij} = \frac{p_i o_{ij}}{d}$$

where p_i is the agreed price, d is the divider of time to get the desired compensation period. To calculate the cost of bandwidth can be formulated as follows:

$$\beta_{ij} = b_{ij} q$$

where β_i is the cost of bandwidth, b_{ij} is the size of data at the time j, and q is the cost per unit of data. β_i is a constant predetermined. Amazon EC2 sell $0.10 for each GB of data transferred. In order to compete with amazon, it should be constant under a set price, q <$

0.000000095 per KB. From equation (7) and (8), can be formulated for the cost of each service provider i at the time j :

$$C_{ij} = \alpha_{ij} + \beta_{ij}$$

The cost for each service provider i is:

$$C_i = \sum_{j=1}^{g} \left(\alpha_{ij} + \beta_{ij} \right)$$

$$C_i = \sum_{j=1}^{g} \left(\frac{p_i o_{ij}}{d} + b_{ij} q \right)$$

The cost for the entire system is:

$$C = \sum_{i=1}^{n} \sum_{j=1}^{g} \left(\frac{p_i o_{ij}}{d} + b_{ij} q \right)$$

This model is called Computational and Communication Cost Model (CCCM).

8. COMPENSATION MODEL

Cost model that has been formulated previously is a model for Internet grids. In this case, the task provider does not have the infrastructure that had been built before, but the resources used for computation is taken from any computer connected to the Internet. Approval of the use of resources is based on the negotiated price using the price model that has been determined. When viewed from the perspective of task provider, then the model is a model to determine the amount of cost incurred by the task provider. When viewed from the perspective of service providers, then the model is a model for determining the amount of compensation to be received by the service provider. By knowing this situation, the compensation that must be paid by the task provider is the cost itself, which can be formulated as follows:

$$K = C$$

$$K = \sum_{i=1}^{n} \sum_{j=1}^{g} \left(\frac{p_i o_{ij}}{d} + b_{ij} q \right)$$

where K is the total compensation. Compensation for a service provider, *i*, is:

$$K_i = \sum_{j=1}^{g} \left(\frac{p_i o_{ij}}{d} + b_{ij} q \right)$$

9. EVALUATION

An algorithm is coded in order to evaluate the model proposed in this paper. Below is the pseudocode for one-to-many bargaining :

```
1.  For CurrentDay As Integer = 0 To
    TOTAL_PROJECT_DAYS
2.  Do
3.    For ServiceProviderIndex As Long = 0
    To TOTAL_SERVICE_PROVIDER – 1
4.      Compare service provider price with
    task provider price
5.      Get minimum service provider price
6.    Next ServiceProviderIndex
7.    Deal with selected service provider
8.  Loop Until
    TotalServiceProvidersChosen.Count >=
    MIN_SERVICE_PROVIDER_PER_DAY OrElse
    TotalServiceProvidersPerDay.Count = 0
9.  Next CurrentDay
```

This algorithm compares the prices between task provider and service providers each day until the project is completed. In accordance with the concept of one-to-many bargaining, service providers offer will be collected first, and then task provider will decide whether the offer is accepted or rejected. In addition, task provider will also make agreements with other service providers if still not meet the required quota, as long as these providers offer still meet the threshold that had been predetermined by task provider.

The calculation of the task provider price can be calculated using the formula :

```
CurrentTaskProviderPrice =
TASK_PROVIDER_START_PRICE + (CurrentDay /
TOTAL_PROJECT_DAYS) *
(TASK_PROVIDER_MAX_BUYING_PRICE - MIN_PRICE)
```

For service providers, the price offered to the task provider can be calculated through the pseudocode :

```
1. For CurrentTaskProviderPrice As Long =
   MIN_PRICE To
   TASK_PROVIDER_MAX_BUYING_PRICE
2.   Dim ExpectedSurplus As Double =
   (TASK_PROVIDER_MAX_BUYING_PRICE -
   CurrentTaskProviderPrice) *
   GetTaskProviderProbability(CurrentTaskPro
   viderPrice, MIN_PRICE,
   TASK_PROVIDER_MAX_BUYING_PRICE, Round)
3.   Compare and take price with max surplus
4. Next CurrentTaskProviderPrice
```

When associated with the formula has been designed in chapter 6 and 7, then:

$$R = x + \left(1 - \left(\frac{f}{h} \right) \right) (y - x)$$

x = TASK_PROVIDER_START_PRICE, f = CurrentDay, h = TOTAL_PROJECT_DAYS, y = TASK_PROVIDER_MAX_BUYING_PRICE, R = CurrentTaskProviderPrice.

$$V = \min(R, \min(w))$$

V = No 4 (Task provider and service providers price comparison), w = No 5 (Get service providers minimum price).

$$E = 1 - \frac{(m - V)}{(z - V)}$$

E = No 7 (Agreements with service providers).

$$E = (m - R)s$$

m = TASK_PROVIDER_MAX_BUYING_PRICE, R = CurrentTaskProviderPrice, s = GetTaskProviderProbability, E = ExpectedSurplus.

Comparison between one-to-one bargaining and one-to-many bargaining is based on some criterias : total service providers is 1000, the maximum price that can be accepted by the task provider is 100, the minimum price acceptable by the service providers is 60, the maximum value of the transaction is 150, the minimum value of the transaction is 10, changes in the probability of rejection of the offer is 0.1, the project has a duration of 30 days, the minimum amount of service providers that should be met per day is 100, and task provider starting price is 50.

Both methods use the same criteria, so that both methods can be compared. These criteria are used to perform variables randomization using uniform distribution. Testing is done from two sides, from the performance and from the stability of one-to-one and one-to-many bargaining.

Research carried out by using Visual Basic in Visual Studio 2008 environment. This research represents the actual conditions, because the variables that will be measured is not about the speed, but the final price from the two methods.

Performance

Testing for performance is intended to determine whether one-to-many bargaining is better than a one-to-one bargaining, or vice versa. Testing is done by comparing the minimum, maximum, average, and the total price. One-to-one bargaining, and one-to-many bargaining will be compared based on the number of available service providers.

From the Figure 7, it is known that the minimum value between the one-to-one bargaining with one-to-many bargaining more or less the same. This condition appears because the nature of these two methods that will immediately take the service providers who offer lower prices than task provider expected price. At the beginning of the graph, it shows a little difference, because one-to-many bargaining has a better method in negotiating by playing service

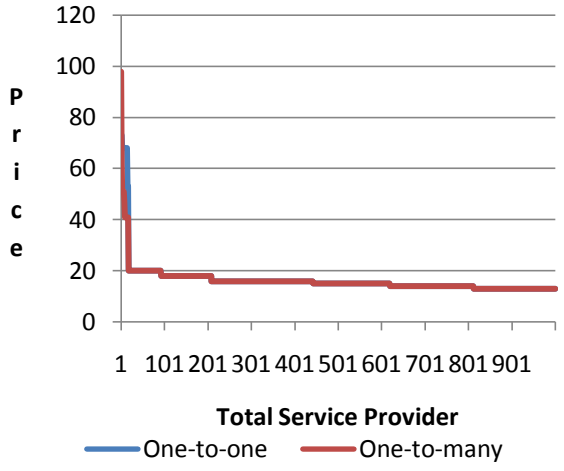


Figure 7. Minimum Price Performance Comparison

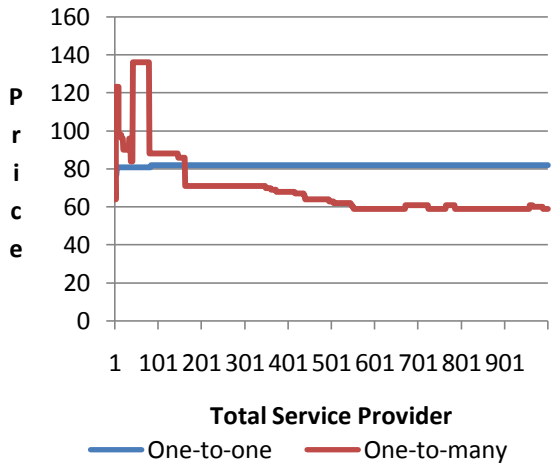


Figure 8. Maximum Price Performance Comparison

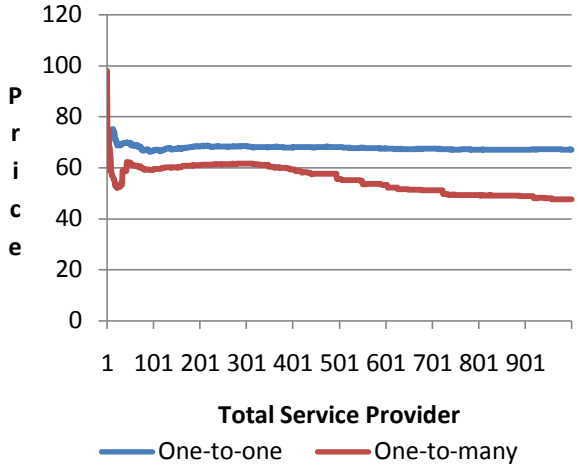


Figure 9. Average Price Performance Comparison

providers time pressures.

One-to-many bargaining tend to have a maximum value better than a one-to-one bargaining on the number of service providers more than 162. This happens because one-to-many bargaining has to meet the minimum number of service providers.

This experiment is done by using the 100 service providers as a minimum value per day, so that one-to-many bargaining will be more stable if the number of providers is over 100. One-to-one bargaining is better because he always negotiate with each provider, but the one-to-one does not have information when the number of service providers are met. This will be very rare in the real world because the number of service providers as much as 1,596,270,108 [8] should be more than enough.

Based on the average price performance, one-to-many bargaining has a lower price than one-to-one bargaining.

Based on the Figure 10, it is shown that when there is only 1 service provider, one-to-one bargaining, and one-to-many

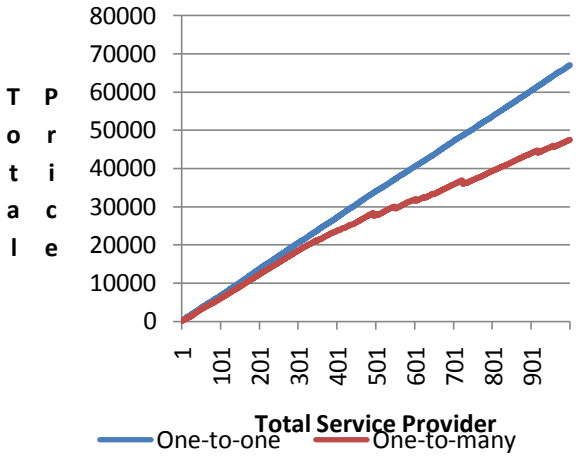


Figure 10. Total Price Performance Comparison

bargaining have the same price, but along with the growing number of service providers, one-to-many bargaining has a lower price than one-to-one bargaining.

Stability

To see the stability of the method, the calculation process is performed 10 times. To get more accurate results, then the results that will be compared is the minimum value of one-to-one bargaining with maximum value of one-to-many bargaining. Figure 11 until 14 are the results of the research has been done. Numbers listed above table shows the number of experiments that are conducted.

Based on the minimum price, the results show that one-to-many bargaining has a stable minimum price, because the price is

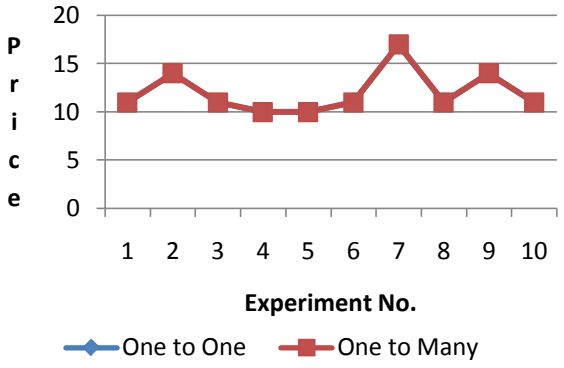


Figure 11. Minimum Price Stability Comparison

consistently same with the price generated by the one-to-one bargaining.

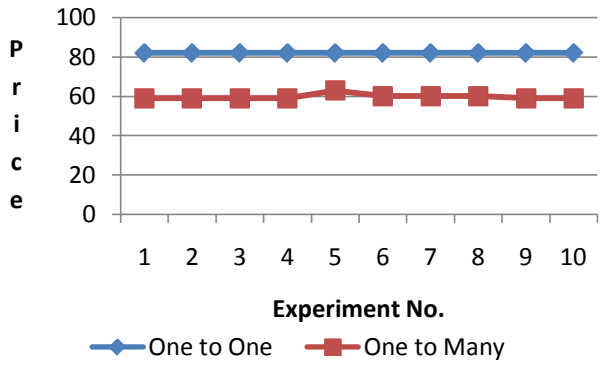


Figure 12. Maximum Price Stability Comparison

One-to-one bargaining and one-to-many bargaining maximum price are equally stable. Based on 10 times randomization, one-to-one bargaining price is around 80, while one-to-many bargaining price is around 60.

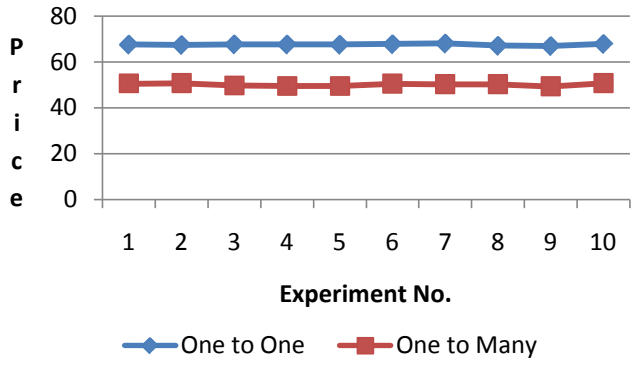


Figure 13. Average Price Stability Comparison

Based on Figure 13, it is known that one-to-one bargaining average price is from 66,979 until 68,059, while one-to-many bargaining average price is from 49,334 until 50,789. By knowing this number, it can be concluded that one-to-one bargaining and one-to-many bargaining have stable maximum value.

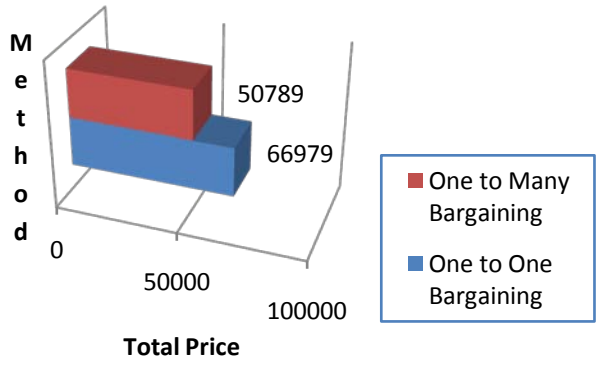


Figure 14. Total Price Stability Comparison

According to the total price, one-to-one bargaining and one-to-many bargaining has a stable price range, 66979-68059 and 49334-50789. If the minimum value of one-to-one bargaining and the maximum value of one-to-many bargaining are compared, it is known that one-to-many bargaining can save 24.17% than using one-to-one bargaining.

10. SIMULATION

The algorithm that is written in Visual Basic code in chapter 9 is implemented on Rocks Clusters 5.1 and Tachyon 0.98.7. Rocks Clusters are used to create a cluster consisting of a head node and several compute nodes. Tachyon is used to simulate the ray tracing on the cluster. Computational and Communication Cost Model (CCCM) will be applied to the tachyon. Bargaining will be conducted before the tachyon is run, and the compensation calculation based on the computation and bandwidth used will be run after the tachyon.

Simulation is done by using a head node and two compute nodes. It represents any number of nodes that will be used, because by observing the number of nodes variable, then the number of nodes doesn't change the result significantly.

The application of one-to-many bargaining on the Rocks Clusters 5.1 is done by using MPICH library written in C. The number of cpu cycles obtained from the linux command, the "`dmesg`". Complete command to get cpu cycle of resource providers in units of MHz is "`dmesg | sed -n 's/.*Detected \\(.*\\) MHz processor./\\1/w /tmp/cpu'`". From that command, the number of cycles will be written in "`/tmp/cpu`", which will be read and sent to the task provider via `MPI_Send` task. The service provider identity is known by using "`gethostname ()`" in "`unistd.h`".

Tachyon 0.98.7 is modified to be used in conjunction with CCCM. Cpu cycles calculation while running ray tracing is need to be added, and bandwidth usage calculation is sent to the task provider after ray tracing task is completed. Cpu cycles monitoring is done by using "`top`". The full command is "`top -b -n1 | grep tachyon | awk '{print $9}' > /tmp/cpuusage`". That command is used to see the cpu usage percentage for a process called tachyon, then write it into a "`/tmp/cpuusage`". This file will be read and added regularly. Interval used is per second. This system does not use "`delay()`" or "`sleep()`", but using "`time()`", because of the time accuracy. "`delay()`" and "`sleep()`" is considered not reliable in handling the delay in the system.

Bandwidth calculation is done when calling MPICH command, `MPI_Send_init`. That command has one parameter, the count, which is used to determine the amount of data to be sent to the task provider. In the tachyon, the amount of bandwidth usage is calculated by using the formula "`totalbandwidth + totalbandwidth = scene-> hres * 3;`".

The total compensation for cpu cycles usage is calculated using the formula "`totalcpuusage * totalcompensation = PricePerSecond;`", while the bandwidth usage compensation is calculated using the formula "`bandwidthcompensation[i] = totalbandwidth[i] / 1000000 * BANDWIDTH_PRICE;`".

Figure 15 is the result screen after CCCM is successfully simulated.

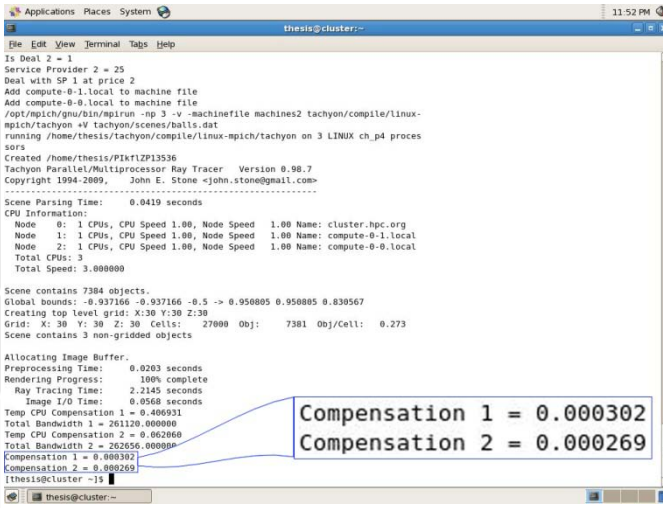


Figure 15. CCCM on Rocks Clusters 5.1

Figure 16 is the ray tracing result.

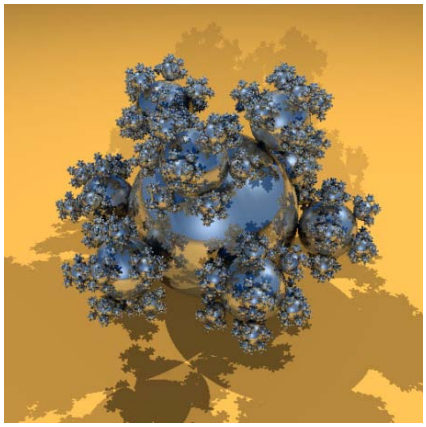


Figure 16. CCCM Ray Tracing Result

11. CONCLUSION

One-to-many bargaining is better than one-to-one bargaining for considering the price offered by service providers. By using one-to-many bargaining, the task provider can save costs 24.17% compared to using one-to-one bargaining. Computational and Communication Cost Model (CCCM) is a cost model that is better than OPTIMAL because this model does not only calculate the computational cost, but also calculate the communication cost. Task provider income will increase by applying one-to-many bargaining and CCCM.

12. REFERENCES

[1] I. Foster, "The Anatomy of The Grid : Enabling Scalable Virtual Organizations," *International Journal of Supercomputer Applications*, 2001.

[2] BBC. (2007) BBC NEWS | Technology | PS3 network enters record books. [Online]. http://news.bbc.co.uk/2/hi/technology/7074547.stm

[3] Stanford. (2009, May) Folding@Home. [Online]. http://fah-web.stanford.edu/cgi-bin/main.py?qtype=osstats
[4] C. Lu. (1997) Applications of Ray Tracing. [Online]. http://www-cs-faculty.stanford.edu/~eroberts/courses/soco/projects/1997-98/ray-tracing/applications.html
[5] E. a. B. P. Afgan, "Computation Cost in Grid Computing Environments," *IEEE 0-7695-2830-9*, 2007.
[6] A. a. A. J. Caracas, "A Pricing Information Service for Grid Computing," *ACM 978-1-59593-944-9*, 2007.
[7] R. Needleman. (2008) Amazon CTO on AWS outage: Like you can do better?. [Online]. http://news.cnet.com/8301-17939_109-9905105-2.html
[8] Miniwatts. (2009) World Internet Usage Statistics News and World Population Stats. [Online]. http://www.internetworldstats.com/stats.htm
[9] J. a. B. R. Carroll. (2009) Online Marketing Guide. [Online]. https://www.paypal.com/en_US/pdf/mkt_guide.pdf
[10] C. Waldspurger, "Spawn: A Distributed Computational Economy," *IEEE*, vol. 18, no. 2, pp. 103-117, 1992.
[11] N. Nisan, "Globally Distributed Computation over The Internet: The POPCORN project," *IEEE*, 1998.
[12] R. Buyya. (2002) Economic-based Distributed Resource Management and Scheduling for Grid Computing. [Online]. http://www.buyya.com/thesis/thesis.pdf
[13] G. Heiser, "Resource Management in The Mungi Single-address-space Operating System," *Proceedings of Australian Computer Science Conference*, 1998.
[14] Y. Amir, "An Opportunity Cost Approach for Job Assignment in a Scalable Computing Cluster," *IEEE*, vol. 11, no. 7, pp. 760-768, 2000.
[15] B. Chun, "Market-based Cluster Resource Management," *Ph.D. Dissertation*, 2001.
[16] SETI@home. (2009, Jul.) SETI@home. [Online]. http://home.ssl.berkeley.edu/
[17] Mojo-Nation. (2001, Jun.) Mojo-Nation. [Online]. http://www.mojonation.net
[18] M. Stonebraker, "An Economic Paradigm for Query Processing and Data Migration in Mariposa," in *Proceedings of 3rd International Conference on Parallel and Distributed Information Systems*, 1994.
[19] D. Grosu, "Load Balancing in Distributed Systems: An Approach Using Cooperative Games," *IEEE*, pp. 501-510, 2002.
[20] P. Ghosh, "A Pricing Strategy for Job Allocation in Mobile Grids using a Non-cooperative Bargaining Theory Framework," *ACM 0743-7315*, 2005.
[21] M. a. R. A. Osborne, "Bargaining and Markets," *Academic Press, Inc*, 1990.
[22] E. Gerding, "Bilateral Bargaining in a One-to-Many Bargaining Setting," *ACM 1-58113-864-4*, 2004.
[23] P. Winoto, "An Extended Alternating-Offers Bargaining Protocol for Automated Negotiation in Multi-agent Systems," *ACM 0-262-51129-0*, 2002.

Mobile RSS Push Using Jabber Protocol

Fajar Baskoro
Department of Informatics, Faculty of Information Technology, Sepuluh Nopember Institute of Technology
fajarbaskoro@gmail.com

Dwi Ardi Irawan
Department of Informatics, Faculty of Information Technology, Sepuluh Nopember Institute of Technology,
penyihirkecil@yahoo.com

ABSTRACT
Rapid development in mobile devices has made its way for mobile phone as an efficient tool to store, process and access any information Vast demand for latest and most accurate information has been a certain distinctive challenge for everyone. This is coused by people who tend not to read the newspaper or using PC to access information through internet while subscribing to some prepaid content is often considered costly.
One of the scenario we can use is by taking the of content that support RSS (Really Simple Syndication). Using Push technology, RSS content can be distributed. Jabber protocol is choosen as a push technology implementation and as a protocol in handling data exchange between client and serve. In this research, the writter is going to develop an mobile RSS Push where user only have to subscribe his RSS channel, while the RSS Push Server will automatically update the RSS channel then distribute it using jabber protocol to subscribed user based on their RSS channel.
The main objective on developing Mobile RSS Push is to provide alternative efficient solution in distribution information and as an actual and efecient information provider.

Keywords
Web Feed, RSS, Jabber

1. INTRODUCTION

Rapid development in mobile devices has made its way for mobile phone as an efficient tool to store, process and access any information. Vast demand for latest and most accurate information has been a certain distinctive challenge for everyone. This is coused by people who tend not to read the newspaper or using PC to access information through internet while subscribing to some prepaid content is often considered costly.
One of the scenarios that can be used is to retrieve the benefit of web content that support RSS. RSS(Really Simple Syndication) is a term in internet technology that refer to the way to syndicate content of a website(web syndication). RSS makes the internet user easily to know the summary or the full web content without a need to visit the website first.
The author's main objective of this research is to develop an client server application where on the client side, author will develop a mobile application to receive news or article that is sent by server, whereas in the server side the author will develop a server application to manage, receive RSS news from RSS address of a website that has been registered by a client and to send lastest news to client using Jabber protocol.

Supported by the GPRS technology, RSS content of a website can be received by the client with a low cost because the data is sent as text and the rate is not use normal sms rates but is calculated based on the number of data (bytes) received by the client.

2. LITERATURE REVIEW

2.1 Push Technology

In term of computer and internet, Push Technology was something fenomenal, beginning when Microsoft made Active Desktop facility in their product, Windows98 in 1997 where In Windows98 desktop screen, we can see live a news web without having access it first, so the information was pushed by the provider to us with no need to pull it first by our own, just like we do when we are browsing or surfing in the internet.
Some observers doubt whether push technology is similar to the Push e-mail and it's not new. Listservs existing methods for distribution of the e-mail messages automatically without the user must make request. Basically the principle is similar to the Listserv's push technology, Push technology allows the distribution of information and figure more than what can be handled by e-mail and real-time system which stores and forwards e-mail's not the design for it.
Push technology describes the types of internet-based communication where the server can send or provide data or information to the client without having client to request first to the server. Unlike the pull technology where the client receives the information or data from the server begin with request to the server first, in other words the client must first pull the data from the server.
On figure1, we can see the illustration of the use of Pull technology. User make a request to the page of a website to know the latest information from it. User are actively making requests to know the latest information from a website's content.

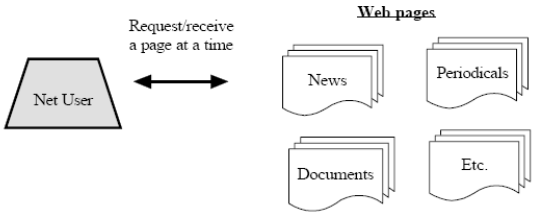


Figure 1. Pull Technology Illustration

Unlike Pull technology, on figure 2, we can see the illustration of Push technology, which server is actively

sending information of a content of a website to the user. User only need to register the website that he want to follow the information inside it. The implementation of Push technology requires software or applications that support both from the client and server, an example of the application that use Push technology is PointCast.

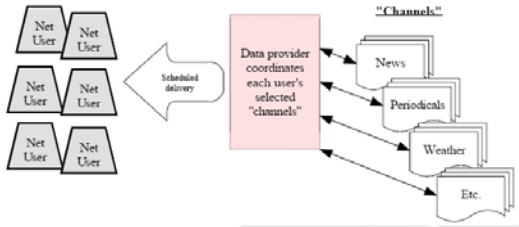


Figure 2. Push Technology Illustration

2.2 Web Feed

Web feed or news feed is a data format that enables us to see and get information on the latest update from the website periodically. The development of web feeds has changed the way publishers publish their content. Web feeds can be a benefit for the publisher to allows them to syndicate content quickly and automatically.
General overview of the visitors who want to know the latest news from a website with the conventional way shown in the figure 3.

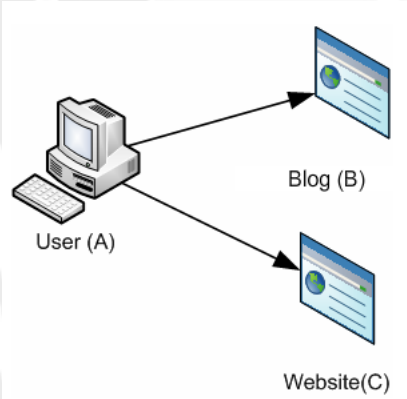


Figure 3. Convensional way to know latest information of a website.

In the figure 3, users is symbolized by user (A) that need to use his browsernya to see if there is new information or news from the blog (B) or website (C). This case it is time consuming and harm users in the use of bandwidth for the volume of data that has been used to open a blog page or website. The direction arrow in the figure 3 shows that the users have to actively visit the blog (B) and the website (C). Compare if we use web feed as shown in the figure 4.

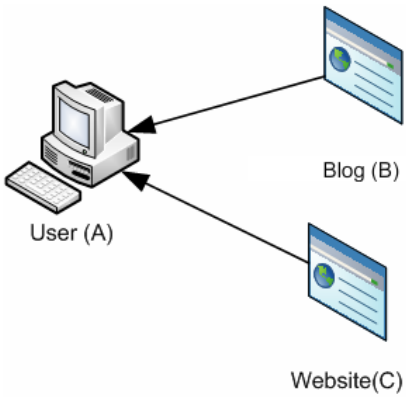


Figure 4. Web feed to know latest information from a website.

The figure 4 users get the latest information from blogs (B) or website (C) without the need to visit it. This is becouse a blog and website have a web feed features. Subscribe to feed of a web site can be done using the application(usually called a feed reader) like web-based application, desktop or mobile. So using feed reader application we will get the latest update of the website or our blog.
In general, the use of web feeds have several advantages, among others :

1. With a web feed, we can choose to view news or information that we want.
2. Save time and bandwidth to get information or news
3. Publisher of a website can publish content quickly and automatically.

Web feeds can be read not only from the computer but it also can be read through other media, such as mobile phone, PDA, smartphone and other devices as long as in that device there is a feed reader application and connect to the Internet.

Figure 5. Web feed in the various hardware

Basically the web feed is a document (usually XML format) that contain the content of a short content with a link to full version. There are 2 kinds of web feed :

1. RSS
2. Atom

Although there are differences between the RSS and Atom, but they basically have the same functionality that is taking

the latest news updates from a website and change it into a format that can be read by a feed reader.
The development of web feed technology alone can we see in the figure 6.

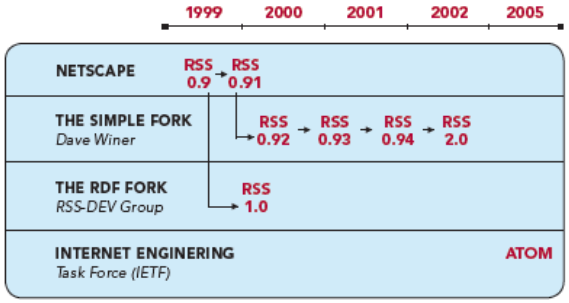


Figure 6. History Diagram of web feed (RSS and Atom)

2.3 RSS 2.0

RSS in one of web feed format that is used to publish an information
RSS is one of the types of web feed formats used to publish information that continues to grow as blogs or news headlines in the standard format. An RSS document provides summary information data (summary), headlines or content of a website full of content with metadata information such as date of publication and author. Use of the word can refer to the RSS format that is much, Really Simple Syndication (RSS 2.0), RDF Site Summary (RSS 1.0 and RSS 0.90), or Rich Site Summary (RSS 0.91).

2.4 Jabber

The presence of IM has become a huge phenomenon on the Internet several years ago. It is marked with a large number of IM users increased. Level of penetration of IM usage is very high due to its ability to facilitate communication in a rapid, and able to rise to the impression of 'no distance'. Ease of use and the level of needs of the user is quite simple are also several of the reason why the IM users increased.
IM is not only used for informal purposes only, but it also used to support research activities and office activities and other industries. In Todays IM there is fragmentation in IM market, some of the IM protocol are already available, although each protocol is still working in the same way. But the presence of Jabber protocol provide a huge effect of development in the IM community because this protocol is open source, has the interesting features and has the capacity to be developed more (extensibility).
Jabber is an XML protocol for the exchange of messages (message) and presence information (presence) of real-time between two users in the Jabber network. There are lot of benefit of using Jabber technology. In the beginning Jabber technology is asynchronous, the IM platform that can be used widely based on the IM network and its function is almost the same as the system of official IM such as AOL Instant Messaging (AIM) and Yahoo Instant Messaging.

2.4.1Jabber Protokol as Messeging Protocol

In many cases the goal of Jabber is to build better IM system that supports the information in real-time presence and messaging. Better IM System in this case are :

1. **Open**, Jabber protocols are free, open, public and easily understood. This makes it easier for anyone to make the implementation of the Jabber without charge a fee for the license.
2. **Standards**, Internet Engeneering Task Force (IETF) has formulated XML protocol as instant messaging and presence technology that was approved under name Extensible Messaging and Presence Protocol, or XMPP.
3. **Proven**, Jabber was first developed by Jaremie Miller in 1998 and now is stable enough, hundreds of developers works using Jabber technology. There are ten thousand of the Jabber server is active now on the Internet and millions of people use Jabber for IM.
4. **Decentralized**, Architecture of the Jabber-like email, so that everyone can make their own jabber server.
5. **Secure**, Jabber servers can be isolated from other networks. In addition, system security using SASL and TLS has been built in the core XMPP specification.
6. **Extensible**, the advantages of using XML namesapces, everyone can build additional functionality over the jabber protocol. To maintain interoperability, common extensions are managed by XMPP Standard Foundation.
7. **Diverse**, many companies and projects using open source jabber protocol to build a real-time application. Developers will not feel "locked" using technology jabber.
8. **Diverse**, many companies and projects using open source jabber protocol to build a real-time application. Developers will not feel "locked" using jabber technology.

2.4.2 Jabber Architecture

Jabber use client server architecture. Jabber client can communicate with Jabber server on their domain Jabber. Jabber domain has advantages in its ability to separate the zone of communication, which is handled by different Jabber server, not like most other IM systems that use a centralized server for all communication zone. In Jabber, message sent by client to the server and then forwarded to the recipient server and delivered to the recipient client.
Data format used for communication on the jabber data format is XML. XML is a standard World Wide Web Consortium for standard data format, for a generic document. All communication in Jabber involving the exchange of jabber package which every packet can be an XML fragmentation. XML fragmentation can be said in the document as a sub document in the stream communication in Jabber

3. SOFTWARE DESIGN

Mobile RSS Push is a software-based on client –server architecture where at the client side is an JavaME application that can be installed and running on the environment of mobile device while at the server side there are two applications, there are : Jabber server and RSS Push Server. Jabber Server role in the traffic data communication between a client using the standard specification jabber protocol. RSS Push Server is a jabber client that role as a information distribution server.

Software architecture of the Mobile RSS Push can be seen in the figure 4.1.

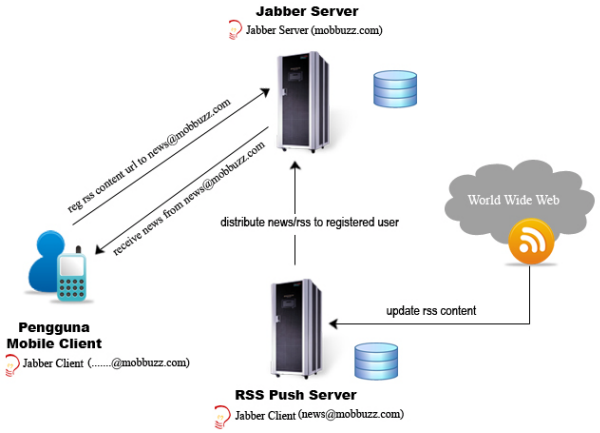


Figure 7. System Architecure

In figure 7 can be seen that basically Mobile RSS Push can be divided into 3 parts : Mobile RSS Push Client, RSS Push Server and Jabber Server. Mobile client users who are registered on the Jabber server send RSS channel registration request to the RSS Push Server. That request will be processed and the RSS channel that has been successfully registered is stored in the database server by RSS Push Server. Next RSS Push Server will update rss feeds regularly for all the RSS channels registered on the database server, then distribute it to all the client according to the RSS channel that has been registered previously.
Scenario at each business system in this application are :

1. Mobile RSS Push Client
   - 1.1. Users can receive and read a message received from the RSS server.
   - 1.2. Users can register their RSS Channel to the RSS Push Server
   - 1.3. Users can categorize their RSS.
   - 1.4. User can foward a message to the other users.
   - 1.5. User can add an ID of other users.

2. RSS Push Server
   - 2.1. Receive an RSS channel registration request from a client
   - 2.2. Processing and updating RSS then parsing it into a format that has been previously determined to be sent to the client.
   - 2.3. Sending updated RSS to the client.

3. Jabber Server
   - 3.1. Jabber Server authorize all the jabber client.
   - 3.2. Jabber Server authenticate all the jabber client.
   - 3.3. Receive stream from the client in the form of jabber protocol standard specification.

3.1 Use case diagram

3.1.1 Mobile RSS Push Client

Use Case Diagram in figure 4.2 descibe that Mobile RSS Push Client has two actors, they are Reader and Jabber servers. Readers can use all the features of Mobile RSS Push Client after connecting to the Jabber Server. Readers can read the latest RSS messages obtained from the Jabber server. Readers can also share their RSS to other readers using RSS Share feature. Reader can register all of the preferred RSS channel so that readers will automatically get the latest RSS from that RSS Channel. Readers can also save the ID of other user. To categorize the RSS Channel, readers can perform categorization on the RSS Channel, eg politics, sports or entertainment.

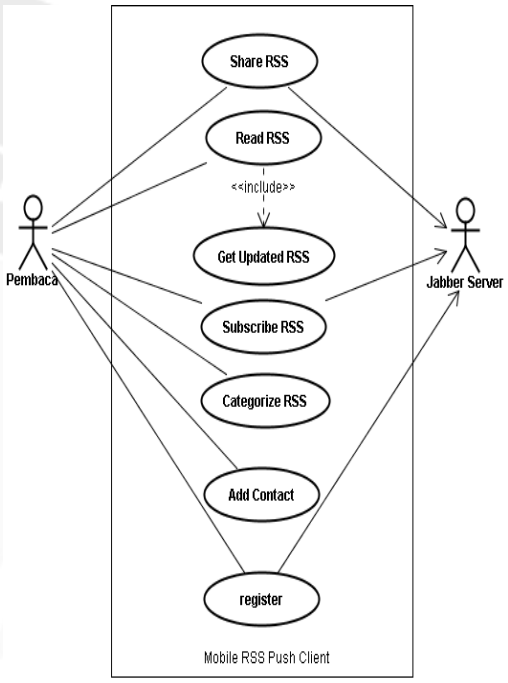


Figure 8. Use Case Diagram of Mobile RSS Push

3.1.2 RSS Push Server

Use Case diagram in the figure 4.3 describes that RSS Push Server has three actors : Jabber Server, Administrator, Web Content. RSS Push Server receives RSS Channel registration request from Mobile RSS Push Client through Jabber server. RSS Push Server automatically updates the RSS channels listed on the database server and send the latest RSS to the Mobile RSS Push client through Jabber client. Jabber Server to act as a mediator in the traffic data transmitted by the RSS Push Server to Mobile RSS Push client. Administrator's job on system is to set the news server which are RSS channel and RSS Channel Subscriber.

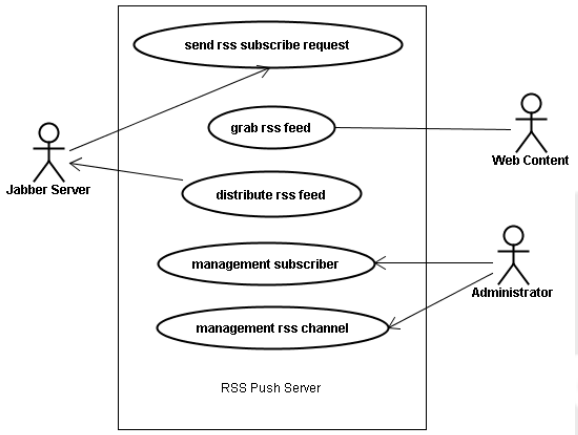


Figure 9. Use Case Diagram ofRSS Push Server

3.1.3 Jabber Server

Use Case Diagram in figure 9 describe that the system has three actors : Jabber Server Jabber, Jabber client and Administrator. Jabber Server is a system that automatically processes send and receive streams from and to one or more Jabber clients that communicate using protocol Jabber. Jabber servers also perform authorization and authentication of the Jabber client. In addition to Jabber server also records all of the stream in the form of message protocol. Jabber client is divided into two kinds which are readers or users Mobile RSS Push Client and RSS Push Server. Readers or users of Mobile RSS Push Client able to register in the system, share and subscribe RSS. RSS Push Server has a role in receiving and registering new RSS. It also has a role to distribute RSS using message protocol of Jabber protocol. Administrator in this system only served in the user management.

3.2 User Interface

This page is displayed when the first application run.

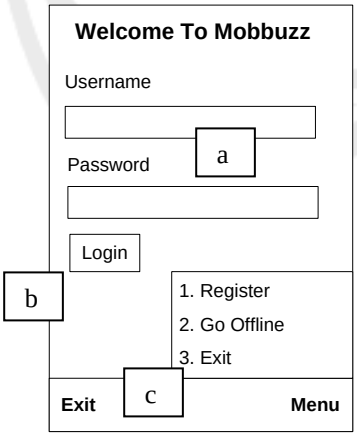


Figure 10. Login Page

Explanation of the figure 10 are:

a. Form login, textField to contain the username and password.
b. The login button.
c. Command, consist of register, go offline and exit.

4. TESTING

For the testing, It will be provided five (5) the website that supports RSS 2.0 technology. The websites are as follows :

- KapanLagi.com
  http://rss.detik.com/index.php/detikcom
- Detik.com
  http://www.kapanlagi.com/feed/
- OkeZone.com - Bola
  http://sindikasi.okezone.com/index.php/bola/RSS2.0
- BBC - Berita Dunia
  http://feeds.bbc.co.uk/indonesian/index.xml
- Kompas.com
  http://www.kompas.com/getrss/nasional

Scenario of this test is to register the RSS URL of five website above using Mobile RSS Push Client and observe whether distribution process of RSS through Jabber server can run smoothly. An incoming message on Mobile RSS Push client is one of the indicators that RSS Push server and Jabber server are running well.

Results test of subscribing RSS Channel can be seen in the figure on 5.1 and 5.2, where in figure 11 is the image of Push Result on Mobile RSS Push and figure 12 is image of a web feed one of website that was created for this test.

Figure 11. Inbox of RSS Message

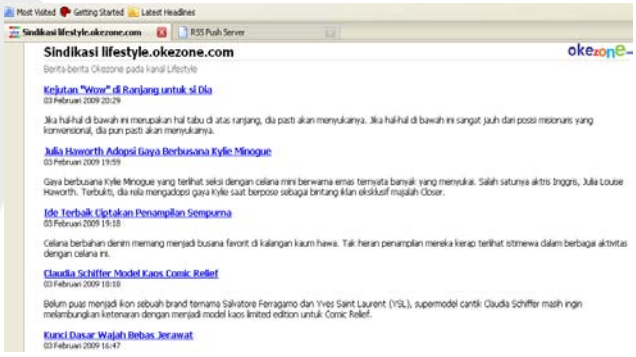


Figure 12. RSS feed pada okezone.com

5. SUMMARY

1. Mobile RSS Push application is designed using the Jabber server that is involved in the RSS distribution and RSS Push Server as a server that taking RSS feeds from the RSS channel that has been registered in the database server. RSS Push has a schedular, a period of time specified by the admin for each update process of RSS channel.
2. Jabber protocol used as a protocol in data transmission between client-server. RSS Push server will send updated RSS Channel to all mobile users who subscribe RSS Channel.
3. RSS Channel registration process is done on the client side. User of mobile RSS Push Client can go to RSS Channel menu by entering the URL of the RSS feeds of a web content that you want to be registered to the server.

6. REFERENCE

[1] http://en.wikipedia.org/wiki/RSS_(file_format) (29 Januari 2008, 11:32)
[2] http://www.w3schools.com/rss/default.asp
[3] http://en.wikipedia.org/wiki/GPRS (01 Februari 2008 12:04)
[4] http://en.wikipedia.org/wiki/Extensible_Messaging_and_Presence_Protocol
[5] (21 Februari 2008, 18:20)
[6] http://www.jabber.org/
[7] http://reader.google.com/
[8] http://java.sun.com/javame/index.jsp

Teacher's Community Building Website to Facilitate Networking and Life-Long Learning

Arlinah Imam Rahadjo
Petra Christian University
Siwalankerto 121-131
Surabaya , Indonesia
+62312983456
arlinah@petra.ac.id

Yulia
Petra Christian University
Siwalankerto 121-131
Surabaya , Indonesia
+62312983451
yulia@petra.ac.id

Silvia Rostianingsih
Petra Christian University
Siwalankerto 121-131
Surabaya , Indonesia
+62312983455
silvia@petra.ac.id

ABSTRACT
Vision and Mission Educational Foundation (YPVM) Teacher's Community Building Website is expected to motivate and empower teachers to keep updating their expertise and professionalism. YPVM, as an educational foundation, has been providing teachers' trainings for years throughout East Java, Indonesia. It has also been aware of the needs of building teachers' community. However no communication media have been available for them yet. By conducting an analysis of the existing system and problems, a web-based Community Building website for teachers is developed. This application is equipped with a Content Management System (CMS) to allow customization of the design and content of the *web*site. The website also includes features on individual spaces for users, discussion forum and online member and seminar registration. An online library and project management tool are also integrated into the system to encourage self-learning and help YPVM to manage its activities. But these two features are not discussed at this paper. The responses of YPVM staff to the system testing are quite satisfactory on the completeness and appropriateness of the application. However further developments should be considered for maximum efforts.

Keywords
Community Building (CB), Content Management System (CMS), Teachers' Training, Web-based Library System, Individual Space, networking

1. INTRODUCTION
Education or human resource development is vital for building the nation. The formal institutions most responsible with this issue are educational institutions. Schools from the level of kindergarten up to high school play important roles in preparing the foundation of education. Teachers, which in this paper also include members of the board and school head-masters, are the main actors in playing their roles as educators.

If we take a look at the data from the Indonesian Department of Education as of March 8, 2008, there have been 243.327 schools in Indonesia (http://npsn.jardiknas.org/cont/ data_statistik/index.php), with 1.293.758 teachers covering those of the elementary up to senior schools (http://nign.jardiknas.org/cont/data_statistik/index.php).

Besides the necessity of going through formal education in teachers' training, quantitatively there have been quite big efforts in informal teachers' trainings in Indonesia for those big numbers of schools and teachers. Seminars and workshops have been set up for the above actors of education by various government or non government institutions to allow them to be kept updated. However the quality of most schools including the actors involved are still in doubt.

The expertise and professionalism of most Indonesian teachers cannot keep up with the growing and changing needs of human resource development to face the national and global challenges. The geographical condition of Indonesia as an archipelago contributes to the problems faced. Most teachers' trainings are held in Java. Limited funds and time from schools and teachers, especially from out of town or outside java, have kept away those educators from updating their knowledge on continuing basis. Specific problems faced by each school or each actor of education from time to time cannot be answered by once and a while seminars and workshops. Limited infrastructure on communication and facilities also contribute to the problems for the educators to share and update knowledge among them or to find references for self development and life-long learning

The building of a community of teachers across the country is an alternative way to facilitate continual sharing of knowledge and experiences among teachers. Building such community is expected to be able to motivate and empower the teachers to keep updating their expertise and professionalism to face the challenges in human resources development of the country. There have been such teachers' communities in Indonesia, though not the online ones. However, again problems of infrastructure on communication leave those existing community not working at their maximum capacity.

At the same time, internet as a powerful means of communication has been growing fast in its utilization in Indonesia. Internet with its capacity of closing the gaps on space and time has the potentiality to bridge the gaps among educators across the country to do networking. Despites of some individual school websites or the government official website for educators providing information on education, up to now there seem to have no online networking media among educators in Indonesia.

Being aware of the needs of developing the expertise and professionalism of teachers, YPVM (Vision and Mission Educational Foundation), a non-profit organization was established in 1999. This organization actively takes part in preparing educators to be servant leaders by holding seminars or workshops for teachers, members of the school boards and head-

masters especially on the subject of *Educational Leadership & Management*. There have been efforts to facilitate networking among those participants for seminars to keep updating their knowledge by giving them opportunities for them to communicate with each other or even with the resource persons of seminars and trainings held through mails and emails on individual basis, but these efforts haven't been working well as expected. They seem to face the problems of coordination and facilitation in motivating the teachers to interact and build a community.

A research has been conducted to try to answer the above problems. The purpose of this reseach is to help YPVM to realize its vision on facilitating networking among the members to motivate and empower teachers to keep updating their expertise and professionalism.

2. BASIC CONCEPTS

The development of this website for teachers is based on the concept of Community Building which, if related to information technology, is called Community Networking (Odazby, 2005). According to Odazby, Community Networking functions as an online media for a community of people with similar objectives, to share in order to develop the community formed. The facilities provided for this media included features such as forums and chatting among members as open communication channels and private chatting between certain number of people and emails as a closed communication channel.(Schuler, n.d). Odazby and Schuler pointed out that community networking which was the online version of community building. It had the power to facilitate communication among members without any limitation of space and had the potentiality to be developed to a large scale of networking up to an international scope, although still only limited to those who had access to internet.

The teacher's community building website is developed using a CMS (Content Management System). CMS itself is software to facilitate the content, design and publishing management of a web (Robertson, 2005). The CMS is used to allow flexibility and customization in managing and developing the website for future purposes. This flexibility prevents the administrator not to depend on the programmers to update the design of the website. It also helps the administrator to keep updating the website not only with new contents but also with various designs.

3. SYSTEM AND PROBLEM ANALYSIS

YPVM, as a non-profit organization, has had no promotion media to introduce its existence to the society at large. YPVM is using invitation letters and registration forms as media of promotion. These basic media of promotion were sent to schools and school foundations, limited to East Java. This condition has prevented YPVM to be widely recognized. Consequently the number of participants was also limited. Up to now there have been 30 schools in Surabaya and 25 Boards outside Surabaya involving around 150 teachers taking part in the seminars and workshops held by YPVM. As it had no system on memberships, YPVM had no information on the exact number of members. Informally teachers who have participated on the seminars or workshops have been automatically regarded as members. Registration for the trainings have also been conducted manually through phones or returning the registration forms.

The facilities as members have been limited to information on trainings conducted monthly. YPVM still had no media for communication among members or between members and YPVM. Consequently no significant networking was created among the members. It was also difficult for YPVM to contact the members if they changed addresses. Not so much feedbacks were gathered from the members to develop the organization. The members communicated individually with other participants or YPVM staff or resource persons through emails.

4. SYSTEM DESIGN

Through an intensive surveys and interviews on user requirements, a teachers' community building website (http://www.ypvm.org) has been designed and developed for YPVM to facilitate networking and life-long learning among teachers. The definition of the teachers is limited to teachers, members of school boards and school head-masters from kindergarten up to high school actors of education who become members of YPVM. MySQL Server was used to build the database, while PHP, MySQL, Javascript dan OS Windows were used to build the application

This Teachers' Community Building Website serves as a website for YPVM, is used as a promotion media to introduce its activities and provides online registration for members and participation to the trainings conducted. It also serves as a media for networking, equipped with modules on Community Building developed using Content Management System. The features of the website are as follows:

a. Community Building
   Community Building is a media provided for its members to communicate and share. Thus a network of people with the same interest is created. Each member can share his/her profile. The members can also send messages or post comments to other members. Alerts to members' birthday are also provided. Adopting some features of Friendster, the community building system allows each member to have a personal space to write his/her profile or to view his/her participation on the activities conducted by YPVM. This system allows the member to be located, contacted and commented by other members. Thus a community will be built among the members. As papers of seminars and trainings conducted by YPVM can be downloaded from this system, members who are not able to participate in those activities can still download and learn from the e-materials.
b. Forum
   Forums for sharing information, ideas and experiences are also provided to facilitate discussion with other members. Members can create their own groups with specific topics for the forum. These forums serve as a media for networking and learning.
   Forums can also facilitate further discussion among the resource persons of seminars and trainings held by YPVM and the members after the seminars or trainings are over.
   To participate in the discussion of certain topics, the members have to join the specific group of the forum they are interested in. The registered members have the rights to suggest topics of discussion in the specific group they are registered. The unregistered members are not allowed to read and participate in the discussion belonged to the specific

group. This forum is expected to document all postings among members, YPVM and the speakers of the seminars held to be access later as resources for self development.

c. Member Registration
YPVM website allows any non-members to do online registration and join as members. Most of the members are school teachers who have access to internet. For those who have no access to internet yet, YPVM provides an opportunity to register at the office of YPVM by providing registration form in hardcopy. The member registration is conducted by letting each member to write his/her personal data manually such as: name, address, job, position, phone, and e-mail. YPVM has a plan to conduct trainings on the use of internet for its members. These trainings are expected to let this research product contribute maximum benefits for the members of YPVM. This feature on online registration is expected to help YPVM to have complete information on members, process and maintain the data of its members efficiently and effectively.

d. Online Seminar Registration and Promotion.
An online seminar registration is also provided for members and non-members who are interested to participate in the trainings held by YPVM. This online registration is expected to attract teachers and people in the educational fields to join YPVM. This feature cut the process of invitation letters as media of promotion, as the members get used to access this website regularly. Regular updated news on the availability of seminars is provided and accessible for those who have access to internet.

e. Book Review
Another feature provided is Book Review on collection posted by YPVM, allowing members to give comments. This feature is equipped by Download Feature used as a media for file sharing among members. Those files are mostly slide presentations and other materials from the seminars held by YPVM.

5. SYSTEM TESTING AND RESULTS

The YPVM *Web*site can be accessed by members as well as non members. However the non-members can only do the online registration and view the front page only. The members have to login to start using the features of this Community Building website. The system was tested by entering the name of Silvia as a username to login the system. The login page can be viewed at Figure 1.

Figure 1. Login Page

After Silvia has been successfully login, Silvia has the right to access the community building and discussion forums. Figure 2 is an interface for Silvia's profile.

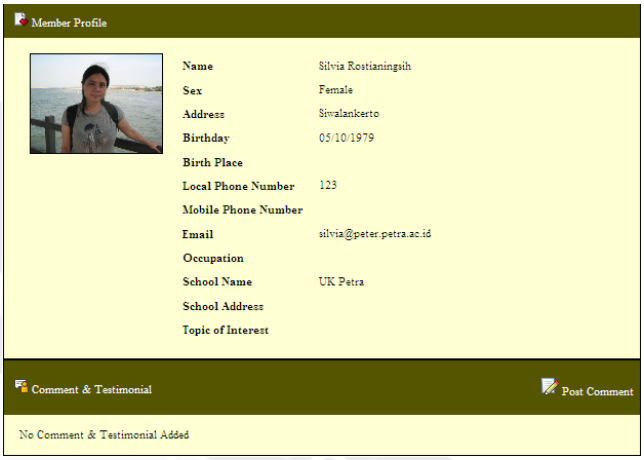


Figure 2. Member Profile

As Silvia enters the system, she is allowed to edit her profile, customize her interface, change password, view other members' profiles, send and receive messages, view birthdays, download files, join discussion forums, post messages and view book reviews. Figure 3 shows the interface of updating profiles of Silvia as a member.

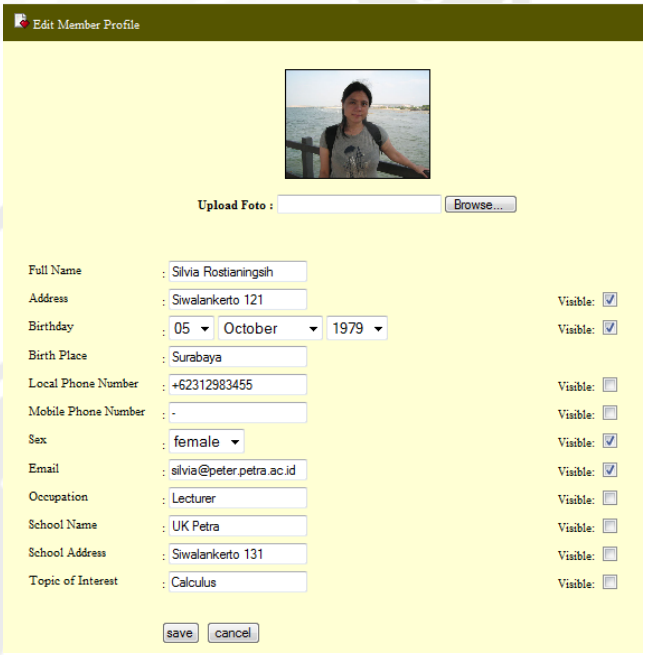


Figure 3. Edit Member Profile

By using CMS, the system provides features for the members to customize the interfaces of their web pages to their own choices. Using the Customized Interface Feature, the members can change the background color, or put pictures at the background. As shown at Figure 4, a member uses pictures of Love as the interface background of his page.

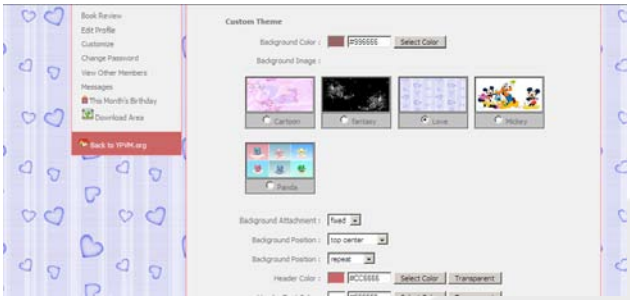


Figure 4. Customized Interface Feature

The user can also view the profiles of other members as viewed at Figure 5. Using this menu, the user can search other members by writing the name of the member searched.

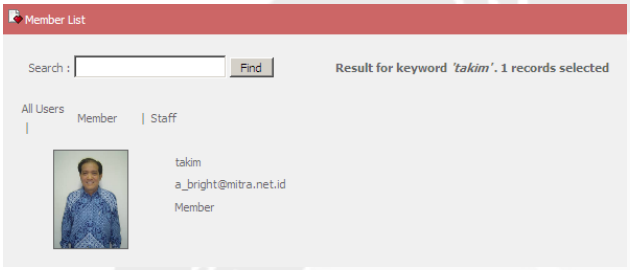


Figure 5. Search Other Member

The User can access the profile of a member by clicking the intended photo. The user can send messages, post comments or invite the intended member to chat as viewed at Figure 6.

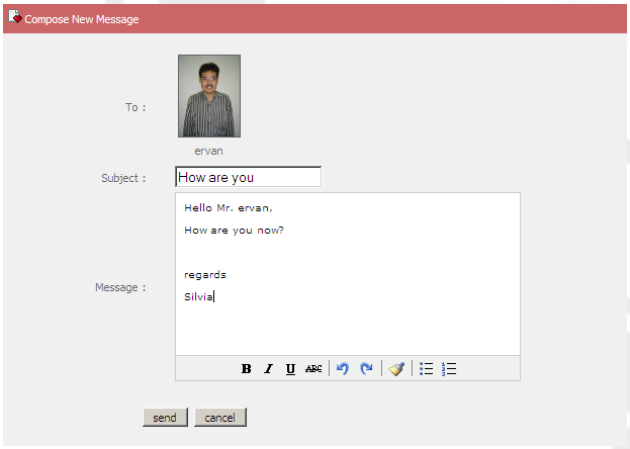


Figure 6. Compose New Message to Other Member

The members have an Inbox and Outbox to keep messages from and to other members.

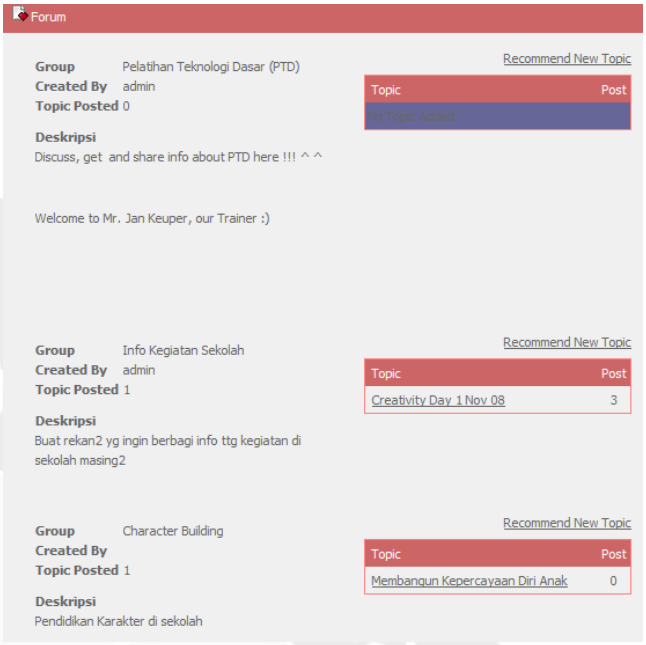


Figure 7. Forum

To join a group discussion, a member has to register to the specific group. Figure 7 shows that Silvia has not yet joined the *"Pelatihan Teknologi Dasar (PTD)"* Group. So Silvia cannot access the discussion topics posted in this specific group. On the other hand Silvia has joined the group on "Info Kegiatan Sekolah". So she can access this group, post new topics or give comments. Figure 7 shows how Silvia as a member of YPVM is making use or the Forum provided at this system.

Another feature provided is Book Review on collection posted by YPVM, allowing them to give comments as viewed at Figure 8.

Figure 8. Book Review

Download Feature is used as a media for file sharing among members as viewed at Figure 9. Those files are mostly slide presentations and other materials from the seminars held by YPVM.

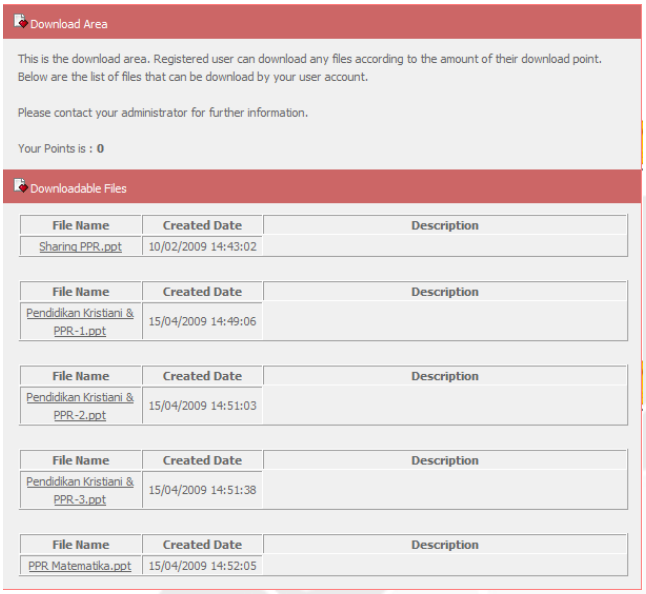


Figure 9. Download Area

6. CONCLUSION

The application developed has been tested by the staff of YPVM. Questionnaires have also been distributed to gather responses towards the appropriateness of the application against the user requirements. The responses of the staff were quite satisfactory towards both the completeness and appropriateness of the features against the requirements (83,33 %).

As this application has just been developed, a test to the real members has not been conducted yet at this time of writing this paper. However this application has also been tested against some representatives of users, including the distribution of questionnaires for them to respond. They also gave quite satisfactory responses on the completeness (75%) and appropriateness of the features against the requirements (80 %), but they gave better responses towards advantages of the website for networking (90 %).

Based on the responses of the respondents and self evaluation on the application, there should be improvement on the features provided. More powerful use of CMS has yet to be designed in allowing more flexibility in expanding menu and template designs, both for administrators as well as for members in managing its own personal space. Inclusion of digital library collection to be provided for the members will give opportunities for members who live far to be able to do the self-learning by access to full-texts. There should also be a facility to upload papers from the resource persons and writers for YPVM magazine. Those papers and online magazine can also be an automatic entry for the digital collection of the library.

Of course the real indicator of success of this website in encouraging networking among members and life-long learning lies in the growing number of members over the years who do the real online networking activities. In due time the indicator of success also lies on the increase quality of members and the schools where they are associated. It still has to be proven yet. There should be another research on this issue.

Another proof of success of this community building website in growing the expertise and professionalism of teachers can also be indicated by the knowledge production of each member in any form of publication. However this objective has to be accommodated by further development of features in the ability to upload, share files among the members and integrate those publication into the digital library collection of the library. There should also be a feature on providing online collaborative writings among teachers.

7. ACKNOWLEDGMENT

The Authors thank to Direktorat Jendral Perguruan Tinggi (DIKTI) Indonesia for giving us a research grant and also thanks to Mustika who have helped in this research.

8. REFERENCES

[1] Departemen Pendidikan Nasional. (2008). Nomor Pokok Sekolah Nasional : Rekap Data. 08 Maret 2008. http://npsn.jardiknas.org/cont/data_statistik/index.php

[2] Departemen Pendidikan Nasional. Direktorat PMPTK DEPDIKNAS (2008). Nomor Unit Pendidik dan Tenaga kependidikan : Data Statistik 2007. 08 Maret 2008. http://nign.jardiknas.org/cont/data_statistik/index.php

[3] Odazby, F. and Frank Odasz. (18 Maret 2008). What is a Community Network? And why you should care? Community Technology Review. 2005. http://www.comtechreview.org/fall-2005/000347.html

[4] Robertson, J. (18 Maret 2008). So, What is a Content Management System? KM Column(03 Juni 2003). http://www.steptwo.com.au/papers/kmc_what/index.html

[5] Schuler, D.(18 Maret 2008). Community Networking: A Network of Sustainable Communities. Heise Zeitschriften Verlag.(n.d.). http://www.heise.de/tp/r4/artikel/8/8024/1.html

Vision and Mission Educational Foundation (YPVM) Web-Based Project Management System

Arlinah Imam Rahadjo
Petra Christian University
Siwalankerto 121-131
Surabaya , Indonesia
+62312983456
arlinah@petra.ac.id

Yulia
Petra Christian University
Siwalankerto 121-131
Surabaya , Indonesia
+62312983451
yulia@petra.ac.id

Edwin
Petra Christian University
Siwalankerto 121-131
Surabaya , Indonesia

ABSTRACT

Vision and Mission Educational Foundation (YPVM) is an educational foundation providing trainings and workshop programs as well as seminars for christian teachers. This foundation was facing problems in integrating and monitoring all activities anytime and from anywhere to be able to produce useful information needed by the foundation to manage and develop further programs.

A project management system application is developed to aid the foundation in executing their programs in its best performance. This application covers access priveleges, jobs arrangement, budgeting arrangement of each activity. The data processing are integratedly arranged. This applicatrion is developed as a web-based application using the PHP and Javascript programming language and MySQL database system.

Resulted from the implementation experimentation, it is concluded that the application has been successful in managing and monitoring every activity conducted. The tasks and budgeting activities can also be presented in Ganttt chart form . Thus the progress of each activity can be monitored clearly.

Keywords

Educational Foundation, Project management, activity

1. INTRODUCTION

Vision and Mission Educational Foundation (YPVM) is an educational foundation trying to prepare Christian young generation to be responsible and productive servant leaders in various fields through its efforts in empowering Christian schools. The efforts are reflected in providing teacher training programs in the form of trainings, workshops and seminars, especially on Educational Leadership and Management based on Christian values. In the implementation process of their activities, there have been some parties involved in planning, organizing, monitoring and evaluating. Good coordination in setting the date, place, job description of each party involved including the budget planning and monitoring the progress of each activity is needed.

Some parties involved are sponsors, the members of the foundation, activity committe, YPVM staffs etc. Some parties such as the sponsors and the members of the foundation are not full-timers. They have their own jobs else where. They also have their own schedules. Problems came up as they have to meet, discuss and make decision in planning and implementing the activities.

Therefore YPVM needs a web-based project management application to help planning and managing each activity without any limitation on space and time. By then all activities can be managed and monitored by all parties involved anytime and from anywhere.

2. BASIC CONCEPTS

2.1 Project Management

According to Schwalbe (2006), Project management is an application of knowledge, skills, tools and techniques in a project activity to meet the needs of the project itself. The success of a project is not only to cover time management, coverage, budgeting and qualities but also has to facilitate all process to achieve the goals.

According to Marion E. Haynes (1993, p.3), "Project Management unify and optimize the resources needed to complete the project successfully". These resources include skills, talent and teamwork effort; facilities, tools and equipment; information, systems and techniques; and money.

Project management has a bond with stakeholders. Stakeholder consists of those involved in or influenced by a project activity. They are sponsors, project teams, supporting staff, customers, users, vendors and those who are related to a project. Every stakeholder has different needs and expectations.

2.2 Project Management Knowledge Area

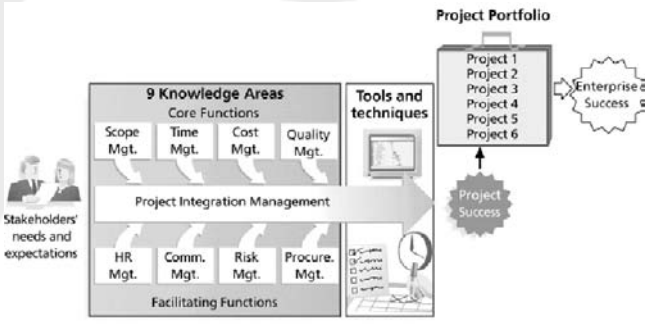


Figure 1. Project Management Knowledge Area (Schwalbe, 2006)

According to Schwalbe (2006), Project management knowledge area describes the key competencies that have to be developed.

Figure 1 shows 9 key knowledge areas from a project management. Four key knowledge areas: scope, time, cost and quality management are considered keys as they play important roles to meet the goal of a project.

Besides those four knowledge areas, there are also other four knowledge areas facilitating project management. They are human resources, communication, risk and procurement management. Those knowledge areas are the process necessary to go through to meet the goal of the project.

The Project integration management is the ninth knowledge area. It is a function to influence and be influenced by the other knowledge areas.

2.3 Schedule Development

Schedule development uses the results of all project time management pre-process to determine the beginning and last stage of the project. There are always several iteration of all project time management process, before a project schedule be completed. The final goal of a schedule development is to set a realistic project scheduling to provide a basis on monitoring the implementation of a project in the time dimension of a project. The main product of this process is a project scheduling, scheduling data model, a baseline scheduling, the intended change and the change for resources needed, activity attributes, project calendar and project management plan.

One of the tools to help the scheduling development process is Gant chart. Gant chart is a tool for providing a standardized format to present information on project scheduling with a list of project activities from the starting up to the end date of a project in the format of a calendar. .

3. SYSTEM ANALYSIS

YPVM is setting up activities such as seminars, trainings, and school grants. So far the implementation of those activities were conducted manually. The preparation, the implementation and evaluation process were not integratedly managed. The board used Microsoft Word and Microsoft Excell to document the activities and budget reports. The non-integrated system created problems for the foundation to manage and retrieve any information on the activities conducted. As the foundation has been progressing. a project management application is needed to manage the needed information for all activities conducted.

4. SYSTEM DESIGN

The system designed for YPVM involves several parties in its implementation such as members of the boards, sponsors, implementation committee, resource persons, and YPVM staff.

Figure 2 shows an example of a flowchart on conducting a seminar activity. This process determines when and where a seminar is set up and who are involved in the implementation of this type of activity. After this process is completed, an implementation committee will take over the responsibility to arrange all requirements needed to implement this activity from the preparation, implementation up to evaluation stage. .

At the preparation stage, the implementation committee is responsible for listing the tasks that have to be conducted by each appointed member of the committee. The system includes the creation of a proposal online. This proposal can be directly viewed by the members of the board. Through this system, they can also write comments on the proposal or give approvals electronically. The approvals will take the activity into implementation stage. .

The implementation stage is a process to fulfill the tasks defined before. The completed task can be reported automatically by each related member of the implementation committee. This process ends as all tasks have been completed.

The evaluation stage is a stage where the evaluation process is conducted. At this stage, the system will check on the gap between planned activities and its realization. The result of the evaluation is presented in reporting formats. The reports will be submitted to the members of the board as proofs on the realization of the activities.

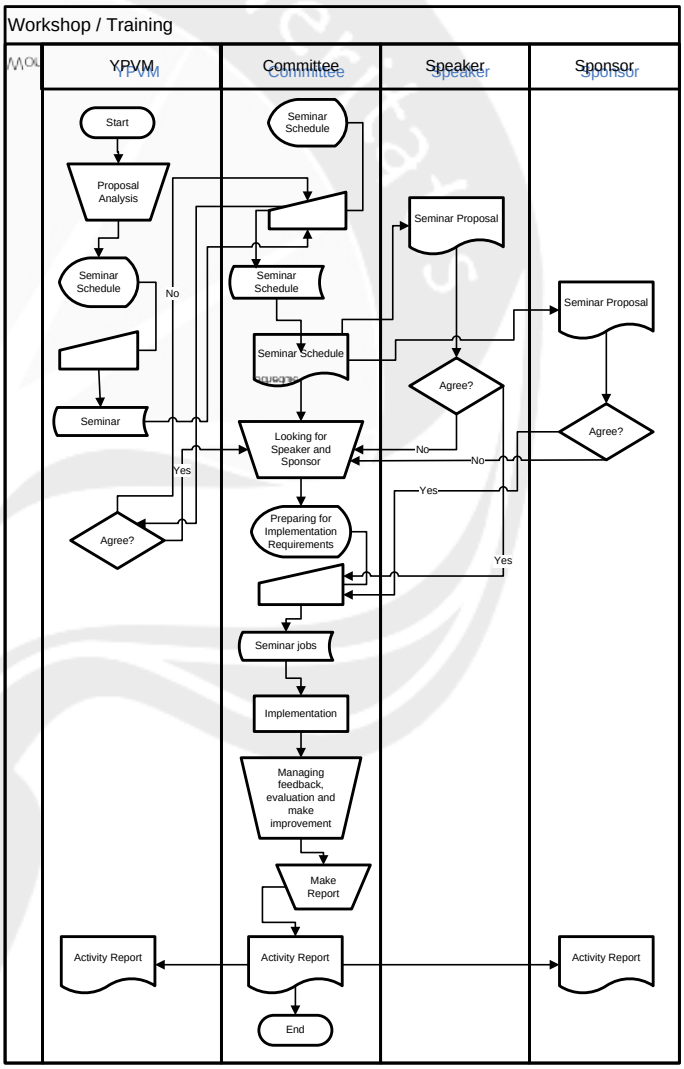


Figure 2. Flowvchart on Seminar Activity

5. SYSTEM EXPERIMENTATION AND RESULTS

Experimentation on conducting a seminar activity is shown on Figure 3 below. It is started by inputting a new event. It is an event on a training program for junior high school teachers.

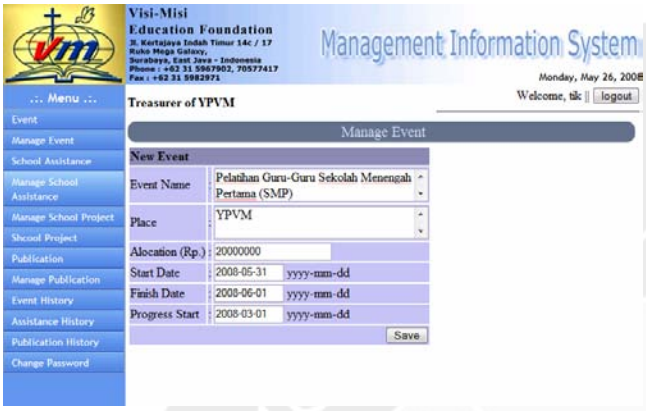


Figure 3. Add New Event

First, an implementation committee is formed. Members are appointed. Each member of the committee will be assigned job descriptions or tasks to be carried out. Each task is divided into three stages: planning, realization and evaluation stages. During this planning stage, a list of tasks, names of committee members and other task predecessors are inputted into the system as seen at Figure 3. Predecessors are determined by inputting the names of the tasks and the detailed information of the tasks.

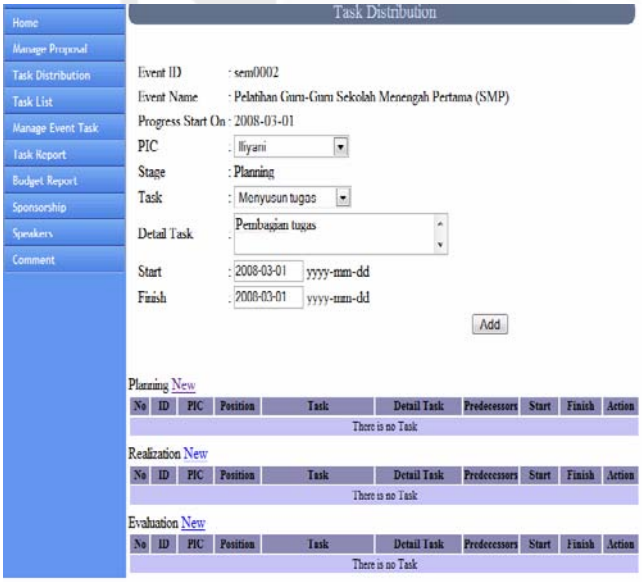


Figure 4. Task Distribution

After the job description has been defined, the next stage is creating a proposal for that activity as seen at Figure 5. Income and expenditure estimation of that event is then inputted.

.

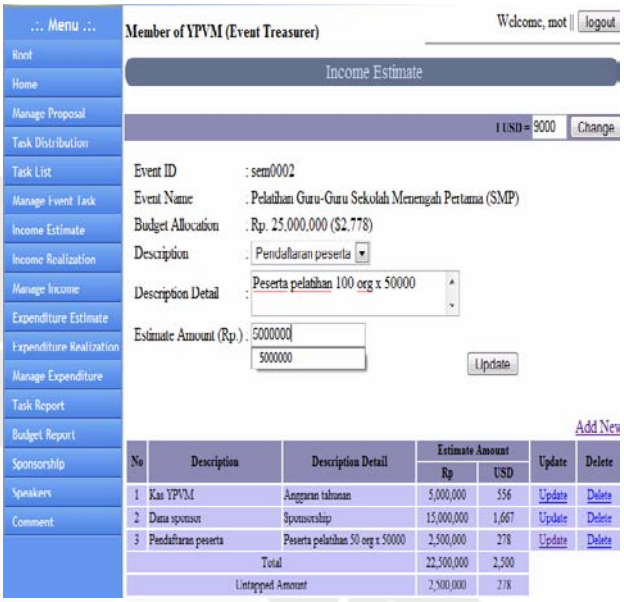


Figure 5. Proposal Inputting

The second stage is the realization stage. At this stage, the implementation of the activities and budgeting is conducted. Figure 6. shows an example of a form on income realization of the event

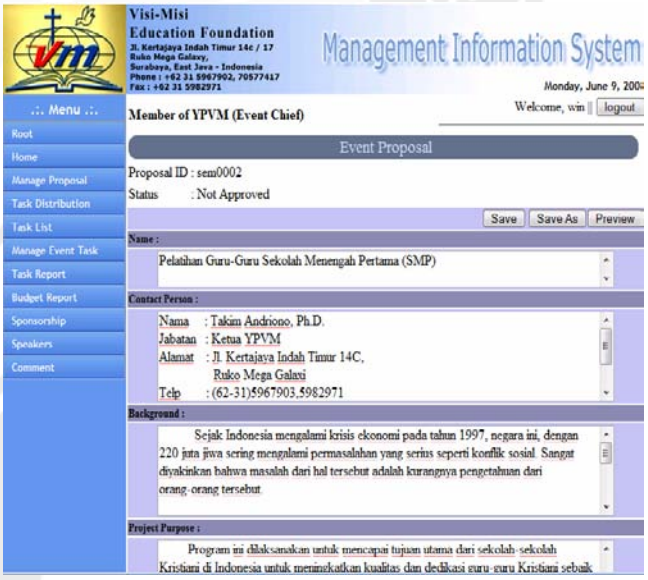


Figure 6. Budget Realization

The third stage is evaluation. There are two types of reports created: reports on tasks carried out and financial reports. Those reports give general and detailed information. Figure 7 is an example of a Gantt chart report on tasks carried out. This Gantt chart presents a report in a graphical form. At this Gant chart, the progress of each stage and the status of each task can be viewed. The blue bar shows the completed tasks. The red bar shows late tasks. The non color one shows that the task has not been conducted yet. There are also milestones of tasks to be conducted.

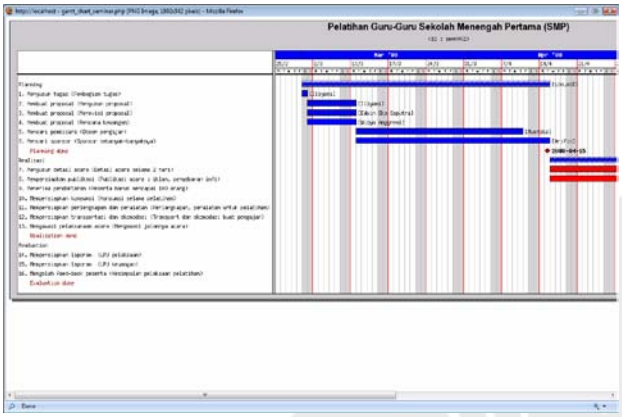


Figure 7. Activity Gantt Chart

The financial report is an overall report on the estimation and realization of income and expenditure. Figure 8 shows an example of an over-budget report in the amount of Rp. 250.000,- from an implementation of an event on a training program for junior high school teachers. The over budget report resulted from the fact that the realization of the expenditure was greater than the realization of the total income.

Budget Report

Event ID : sem0002
Event Name : Pelatihan Guru-Guru Sekolah Menengah Pertama (SMP)
Budget Allocation : Rp. 25,000,000 ($2,778)

Details

Income

No	Description	Description Detail	Estimate Amount		Realization Amount	
			Rp	USD	Rp	USD
1	Kas YPVM	Anggaran tahunan	5,000,000	556	5,000,000	556
2	Dana sponsor	Sponsorship	15,000,000	1,667	15,000,000	1,667
3	Pendaftaran peserta	Peserta pelatihan 100 org x 50000	5,000,000	556	6,000,000	667
Total			25,000,000	1,113	26,000,000	1,224
Gain					1,000,000	111

Expenditure

No	Description	Description Detail	Estimate Amount		Realization Amount	
			Rp	USD	Rp	USD
1	Publikasi	Pembuatan iklan, penyebaran infor	3,000,000	333	3,000,000	333

Figure 8. Financial Report

This system has been experimented on 10 users. The results can be viewed at Table 1, using values from 1 (bad) – 5 (exceptional)

Tabel 1. User Experimentation

No	Questions	No of Respondent				
		1	2	3	4	5
1	Does this appliation match the system of the institution?	0	0	3	5	2
2	How good are the features provided by this application?	0	0	1	9	0
3	Do the reports procuced by the system match the needs of the institution?	0	0	2	7	1
4	How good is the overall user interface of this application?	0	0	1	7	2
5	How good is the overall performance of this application.	0	0	6	3	1

Based on the results of user experimentation as viewed at Table.1, 70% of the users stated that this application had matched the business process of the institution. 90% of the users said that the features provided had been good. 80% of the users acknowledged that the reports produced had matched the needs of the institution. 90% of the users said that the user interface had been good, though the overall performance of the application is only average.

6. CONCLUSION

- The budgeting estimation and realization system for each activity can help the foundation in monitoring the utilization of budgets.
- The availability of a clear time-scheduling on every job description helps the implementation process of each activity carried out as planned. This scheduling system is supported by an information system on the tasks that have to be completed. .
- Based on the results of the questionnaires distributed to the staff and non staff of YPVM, it is concluded that overall the developed application has quite met the requirements.

7. REFERENCES

[1] Haynes, Marion E. 1993. Manajemen Proyek. Jakarta : Binarupa Aksara.

[2] Schwalbe, Kathy., Ph.D., PMP. 2006. Information technology project management. (4th Edition). Canada : Thomson Course Technology

Web Based School Administration Information System on LOGOS School

Djoni Haryadi Setiabudi
Petra Cristian University
Siwalankerto 121-131
Surabaya, Indonesia
+62318439040
djonihs@petra.ac.id

Ibnu Gunawan
Petra Cristian University
Siwalankerto 121-131
Surabaya, Indonesia
+62318439040
ibnu@petra.ac.id

Handoko Agung Fuandy
Petra Cristian University
Siwalankerto 121-131
Surabaya, Indonesia

ABSTRACT

For years, LOGOS Reformed Evangelical Education does not have a centralized and organized academic information system. The data is being kept in archives and this situation complicates the school to control the school's work and quality, such as student's grades data that are still kept in paper archives, unstructured student's data, manually written cash flow transaction and others.

Based on that problems, the solution is to make a design and implementation of web-based school administration information system on LOGOS School. The system will facilitate the arrangement and data storage for the school administration information system's data, and the school's cash transaction system. This web based school administration information system is using internet as its media, and constructed using PHP v5.0.5, *PHP Designer 2007* and *MySQL 5.0*.

From the implementation and application setting, it can be concluded that the system is useful enough to be used as school administration information system. However, it also can be concluded that there is a lack of interface design in the report forms.

Keywords

School Administration Information System, Web-based

1. INTRODUCTION

The school administration and finance at Logos School so far not been computerized and is only using Microsoft Excel software. For school administration system, do not use the database. Financial systems, is very inefficient because they still use the paper and the length of time required to calculate the total of revenue and expenditure.

Therefore they need a good information system and computerized work processes to support this school. Because the users of this system is not only teachers and employees in school, but also the parents who need access to their children's grade from home, so Logos School needed a web-based application.

2. BASIC CONCEPTS

2.1 Administration

Administration has two important aspects, namely: [2]

- Administration of a specific function to control, mobilize, develop, and directs an organization that runs by an administrator who is assisted by his staff.
- Administration is a joint implementation process or the process of collaboration, between groups of people in particular to achieve a particular goal that has defined and preplanned. Cooperation among people took place in and through the organization.
- Data processing system of students, teachers and other school activities are the main functions in an activity of the school administration. Provide better information services to parties in that institution and other related parties outside institution[1]

2.2 Student Information System

A student information system (SIS) is a software application for educational establishments to manage student data. Student information systems provide capabilities for entering student test and other assessment scores through an electronic grade book, building student schedules, tracking student attendance, and managing many other student-related data needs in a school, college or university. Also known as student information management system (SIMS, SIM), student records system (SRS), student management system (SMS) or school management system (SMS).

These systems vary in size, scope and capability, from packages that are implemented in relatively small organizations to cover student records alone, to enterprise-wide solutions that aim to cover most aspects of running large multi-campus organizations with significant local responsibility. Many systems can be scaled to different levels of functionality by purchasing add-on "modules" and can typically be configured by their home institutions to meet local needs. See Figure 1.

Until recently, the common functions of a student records system are to support the maintenance of personal and study information relating to:

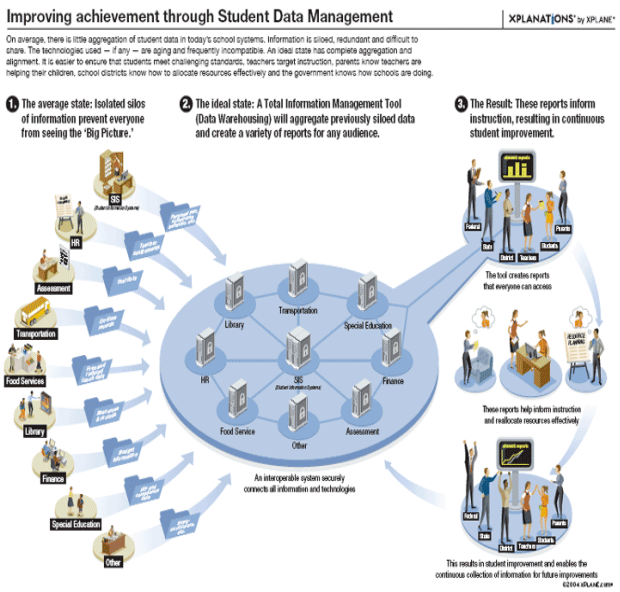


Figure 1. Diagram showing the importance and result of well thought out Student Data Management. [3]

- Handling inquiries from prospective students
- Handling the admissions process
- Enrolling new students and storing teaching option choices
- Automatically creating class & teacher schedules
- Handling records of examinations, assessments, marks and grades and academic progression
- Maintaining records of absences and attendance
- Recording communications with students
- Maintaining discipline records
- Providing statistical reports
- Maintenance boarding house details
- Communicating student details to parents through a parent portal
- Special Education / Individual Education Plan (IEP) services
- Human resources services
- Accounting and budgeting services
- Student health records

In larger enterprise solutions that have student data at their core, further functions include financial aid management and more may be customized by the developer.

3. SYSTEM ANALYSIS

Overall, the manual processes that occurred in the Logos School are as follows:

Biodata collection consists of two parts, the personal data of employees, and the biodata of students. In current systems if the user wants to view the data, then the user must open an existing file on the biodata administration.

Documenting employees made through the curiculum vitae proposed by the employees. (Figure 2). The process begins when the prospective employee filed an application to the Logos School. After passing the selection process by the foundation, if the applicant is administration employee, the applicant will be accepted immediately, if the applicant was a teacher then the applicant will be given special training on how to teach students that are based on Christian values. Once training is complete and the foundation approved, the prospective teachers become permanent teachers.

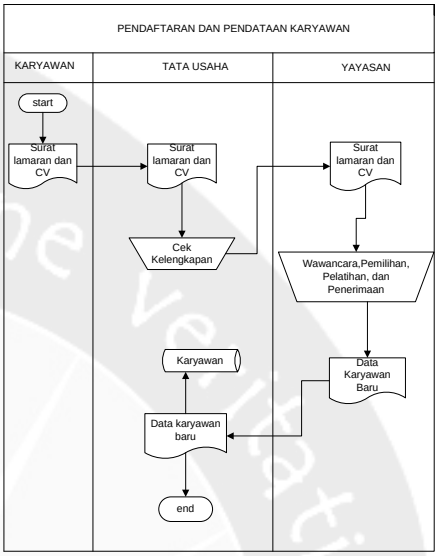


Figure 2. Employee data collection

Admission of the students carried out through a new student registration form. Students who have been selected and accepted, the data stored by the administration. Student data collection process is as shown in Figure 3.

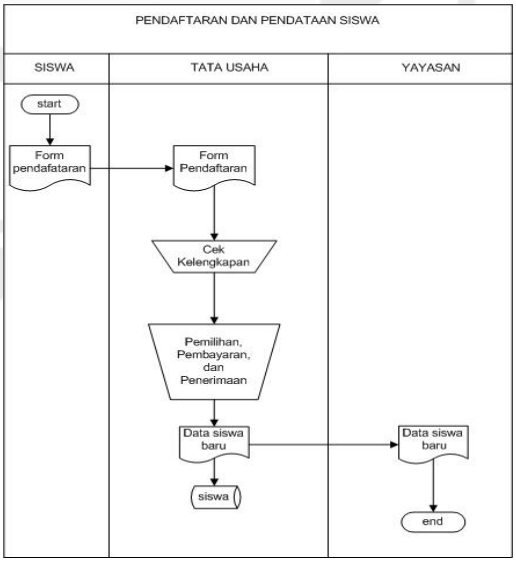


Figure 3. Student data collection

Curriculum distribution is handled by the foundation. The curriculum then submitted to the administration to be stored and delivered to the teachers who teach, then come to the stage of the scheduling details of the curriculum.

System assessment is conducted by each teacher. Teachers have the right to determine the assessment to their students. The grades that have been calculated, then submitted to the administration and to

students or parents of students as a notification. Every day, teachers make a report evaluating each student, which will then be submitted to the administration. Daily values will be accumulated into a report values per week, and then from those reports will be compiled as report cards on each term. This process, as shown in Figure 4.

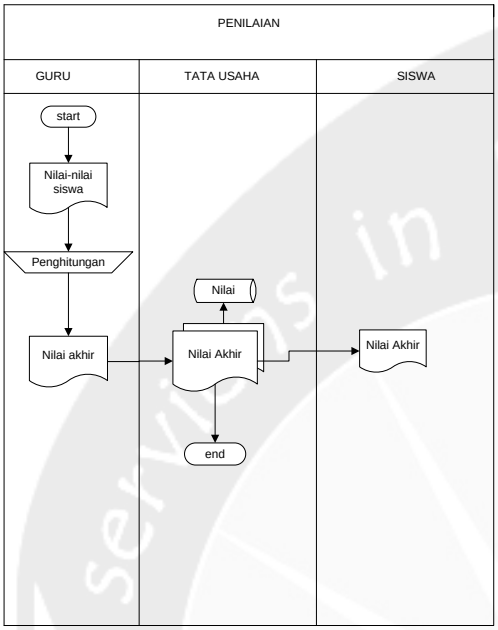


Figure 4. Grading system

4. SYSTEM DESIGN

In the Data Flow Diagram (DFD) on Figure 5 can be seen that The Logos School System consists of four main systems of registration systems, learning systems, assessment systems, and administration systems. Admission system begins when the parents enrolled the students by filling out the student data on the registration form, then this data will be stored into the student and parents database on student registration process.

Teaching and learning system begins when the foundation curriculum materials distributed to the administrator, then the administrator will distribute the schedule of lessons in accordance with curriculum materials. Teachers will be notified of teaching schedules and the biodata of students in classes he taught. Students can view the schedule via the internet. Once that happens everyday educational process which is accompanied with several announcements that can be viewed by students via the internet.

Assessment system starting from the day-to-day assessment conducted by teachers against students. Teachers will fill the daily grades to the database, then from the day-to-day grades, it will be the process that will produce a report calculating the value (on average) each week. The process of making the report card each term is also based on the average weekly value of the student. Results from this report will be deposited into a special database that stores data for each term to the students report card. System Administration covers the development of curricula by the foundation, creation of class schedules and lessons, school revenue, school spending, school announcements, and delivery of student report cards to the foundation.

Application to be designed is a web-based application software it could take several supporters among others:

• Apache 2.0.55, was used as a web server

• PHP version 5.0.5, is used to execute PHP scripts contained in files with extension. PHP (included in installation package XAMPP 1.5.0-pl1).

• MYSQL 5.0, is used as a database

• PHP Designer 2007, used to create interfaces and programming

• phpMyAdmin 2.6.4-pl3, used as a program assistant MySQL database administration.

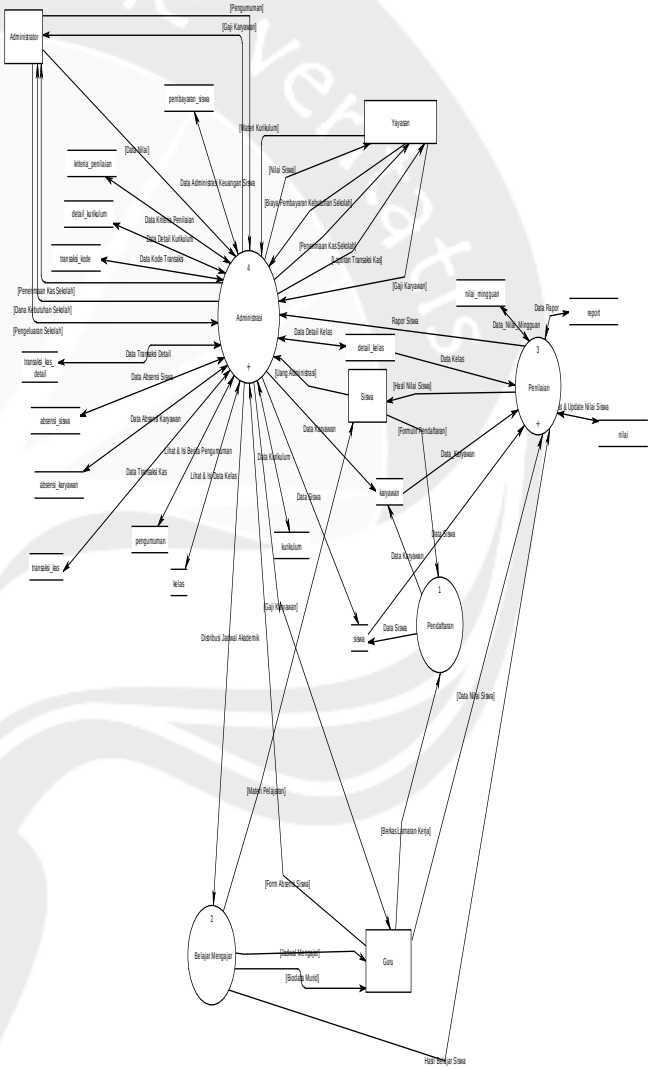


Figure 5. DFD of Administration Information System on LOGOS School

5. RESULTS AND SYSTEM TESTING

Biodata

At first the system testing begins with entering biodata of students. Administrator entering data pertaining to students from the registration form. After filling the data, the status of students is not registered automatically because students are still considered to apply only. After completing the financial administration of building fund, the student status is converted into an active student, at the same time the system will automatically give the Registration Number to the students. (Figure 6).

Figure 6. Form filling new student biodata

For the biodata for teachers and employees, administrators enter the data in relation to teachers or employees concerned, the important thing to consider is the position to the teacher / employee. Then the system will automatically issue the Employee Number of teachers or employees. Data marked with an asterisk (*) must be filled out correctly. Form for new teachers and employees can be seen in Figure 7.

Figure 7. Form new teachers and employees

Curriculum

This feature serves to store data for each level of academic curriculum and each of the existing term. One academic level is run for one year, each year divided into four terms, and each term has a period of ten weeks. Changes can only be done by the administrator. Form to view a list of curriculum as shown in Figure 8.

Figure 8. Form to view a list of curriculum

Grading System

Figure 9. is a form to enter the daily grades of students. Only teachers can entering the grades. Other users can only view the student grades. There are two types of grades, firstly is the value in number (optional), secondly is the descriptions from the teachers concerned about students related with the theme of the criteria. Teachers have the rights to make changes to the value already entered.

Figure 9. Form to enter the daily grades of the students

Weekly grade can be obtained by dividing the total grades of a student in one week with a total attendance of the student. Form of the weekly reports can be seen in Figure 10.

Final report is a result of the division of the total weekly grades of students with a number of weeks during one term, but for the description, teachers need to fill manually. After filling is completed, the Submit button is pressed, then the outcomes of these assessments will be saved and the system will display the

relevant results of the student report card complete with the progress of the students accompanied with graphs, in accordance with their respective assessment criteria. The results of report cards can be seen in Figure 11.

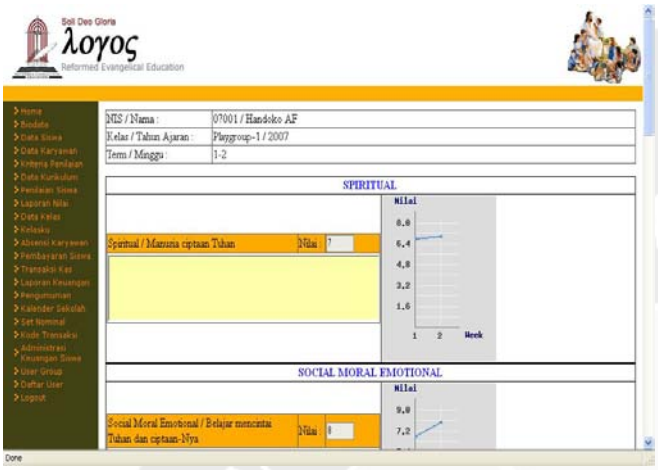


Figure 10. Form to view weekly grades

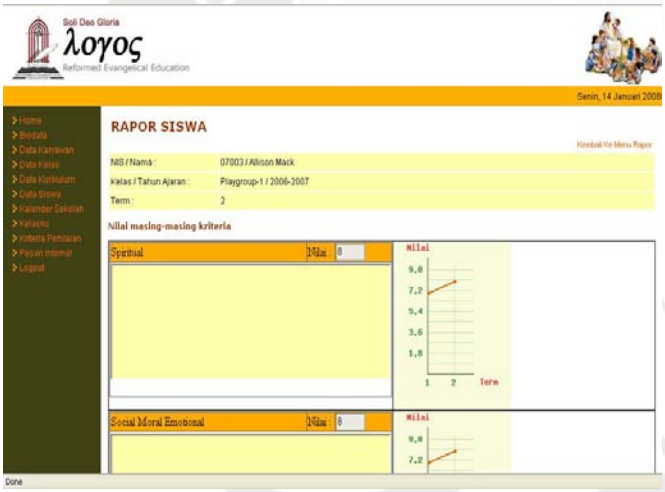


Figure 11. Form of the final grades result

Financial administration

This feature has two parts, namely student payment menu that serves to record cash receipts from students, and student financial administration menu that serves as a processor receiving reports from students. This menu is only accessible by administrators and the foundation. Administrators act as data processors, while the foundation can only view the existing data. For cash receipts, the administrator must enter financial data and after being stored, the system will display the detail transactions that occurred (Figure 12).

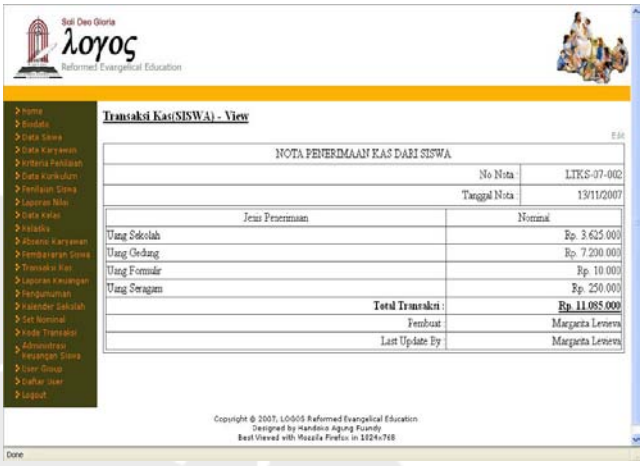


Figure 12. Form to view payment of the student

For Students Finance Administration, the administrator must enter the month and year of the transaction receipts from students, then pressed the button. The system will automatically calculate in detail and display the transaction report that occurred in that month. Results ot the reports can be seen in the Figure 13.

Figure 13. The students financial administration report

6. USER EVALUATION

Evaluation of this system is done by analyzing the results of questionnaires from some users who carried out tests on this system. Users include the following: Administrator, Teacher, Parents, General User.

Assessment using a scale of 1 to 10, where the value of 1 represents lowest value while the value of 10 represents the highest value.

The evaluation criteria to be used on Table 1 are :

1. Feasibility of the software
2. The accuracy of the data presented
3. Speed to presents information
4. Ease of using the software
5. User interfaces appearance

Table 1. User Evaluation

Criteria / User	1	2	3	4	5
Immanuel S	8	9	8	8	7
Pitra	8	7	8	7	7
Tika	8	9	9	9	7
Stefanie	8	7	7	7	6
Sylvia	8	8	8	7	7
Total	40	40	40	38	34

Broadly speaking, through a user evaluation results showed that in terms of appearance and ease of the program scored to 72%, in terms of accuracy and speed of the program scored to 80% and in terms of feasibility of the program scored to 80%. So we can conclude that the system is feasible to use as the school administration

7. CONCLUSIONS

- This LOGOS School Administration Information System has been able to run the administrative information system, student data collection system, data collection system of curriculum, assessment system includes a chart of students progress reports and the financial system.
- After evaluating the implementation of this application and analyze the results of questionnaires relating to the display, can be concluded that the information presented, especially in the form of the report is still less attractive.
- From the results of questionnaires about the software showed that in terms of appearance and ease of the program scored to 72%, in terms of accuracy and speed of the program scored to 80% and in terms of feasibility of the program scored to 80%. So we can conclude that the school administration system is feasible to use to process data streams at Logos School.

8. REFERENCES

[1] School Administration Online. (2007.). Januari 9, 2008. http://www.media.diknas.go.id/media/document/ 4380.pdf

[2] Atmosudirdjo, S. Prajudi. (1980). Administrasi dan manajemen umum volume II (8th ed). Jakarta : Ghalia Indonesia.

[3] Student Information System, Retrieved 2009-03-19 http://en.wikipedia.org/wiki/Student_information_system

Data Visualization of Modulated Laser Beam Communication System

Zin May Aye
University of Computer Studies,Yangon
U.C.S.Y
Yangon, Myanmar
zinmay110@gmail.com

ABSTRACT

The optical communication is one of the important systems in the field of communication. To communicate between two locations, there need to use an adjustable light source to send photons or light wave to the light sensitive photodetector which can detect or collect the light waves. For the first consideration of the system the laser beam transmitter is needed to be constructed. Then the second part is to implement the laser beam receiver. Modulated laser beam transmission is applied in communication. Semiconductor laser diode is used as the light source. The receiver is tuned to receive the modulated laser and rejecting the unmodulated wave. The third part of the system is interfacing. It is to be constructed to convert analog signal to be proceed for computer for further processing. This analog to digital converter hardware circuit is then interfaced with computer line printer port for the purpose of getting information of the incoming signal from the hardware circuit. The communication range is mainly depend on the intensity of the laser diode, transmissitivity and the sensitivity of the photodector.

Keywords

opticalcommunication;photodetector;laser; A/D converter

1. INTRODUCTION

Laser beam communication can demonstrate the elementary concept of modern communication system. In this process laser beam is needed to be modulated. Such modulation adds information (message signal) and carried to be transmitted to a distant location. Most of the method chosen for optical transmission system is a simple on/off light pulses that carry information. The transmitter unit comprises of buffer amplifier and voltage controlled oscillator circuits and the current driver circuit for the laser.Basically the optical communication system works via waves , either sound or EM wave (Radio , Infrared , Laser). The distance between the light source and the object of detection is doubled , the intensity of the light shining on it becomes one-fourth of what it was. This is the reason why the technology to send light from one point to another was very inefficient. Whereas the light from other sources spreads out, but laser is extremely directional. The potential for transferring information using laser was great. The basic process of the communication system is sending information from one place to another. It consists of an information source and destination connected by a communication media,hat transfer messages from transmitter to receiver. The communication channel is the path or medium for electrical or electromagnetic transmission between the transmitter and the receiver. This may be guided or non-guided channels such as radio wave, microwave and the laser beam. The potential for transferring information by laser is very efficient because of the properties of the laser beam: directional, coherence, and monochromaticity [7].

At the receiving end,it is required that a corresponding back-mapping is done to reconvert the signal into the original message. This process is referred to as the demodulation. The associated module with this process is the receiver. In the entire process the laser transmitter, laser beam detector, phase locked loop operation and analog to digital conversion are the main performance in implementing the system.The complete functional block diagram describes overall circuitry processing of the implementing system.

The basic function of this system delineates the transmission of audio signal carried by modulated laserbeam and place information to a point where the laser focuses on the photodetector. The message siganl must be extracted from the laser beam where sound signal is recovered from the laser for listening. In this modulated laser based communication system, the receiving audio signal is needed to be converted in the form of do that it can be interfaced and the digitized signal is visualized on the PC's screen provided by VB software program. The range of the laser based communication is fundamentally depending on the power the laser beam used in the transmitter and the sensitivity of the optoelectronic sensor.

In this paper, audio transmission based on the laser beam is described with the use of laser pointer as a source and the phototransistor is applied for the laser beam detector. The first part introduced the introduction of the system. The second part delineates background of laser communication and laser pointer characteristics. The third part includes design and implementation of the system. The final part described the conclusion of this paper.

2. BACKGROUND

2.1 Laser Communication

Since the basic process of the communication system is sending information from one place to another.It consistsof an information source and destination connected bya communication media,that transfer messages from transmitter to receiver.The communication channel is the path or medium for electricalorelectromagnetic transmission between the transmitterand the receiver. This may be guided or non-guidedchannels such as radio wave, microwave and the laser beam. When the communication medium is air, there is little impedance to prevent a particular pulse shape from reaching its destination in the same configuration as whenit is transmitted. The potential for transferring informationby laser is very efficient becauseof the propertiesof the laser beam: directionality, coherence, and monochromaticity [7].
In this case, the signal must be attached to or superimposed on other voltages at frequencies that move easier in the transmission medium. The process of attaching signals to other easier to propagate signal (carrier) is called modulation. In design and implementation of this system the transmitter unit comprises of

inverting amplifier for pre-emphasis and voltage controlled oscillator circuits and the current driver circuit for the laser. The semiconductor laser diode from the laser pointer module having the output power of less than 3 milliwatt is used as the source of laser emitter.

2.1 Laser Diode

Most of the low power (< 50mW) laser diode are designed to be running in a continuous mode to be rapidly modulated. In order to understand the basic workings of a laser, it is not necessary to understand advanced mathematics or physics. All lasers contain the same basic elements: an amplified medium, feedback, and a power supply. The output of the laser diode is a function of the current flowing across the active junction region. The spot size of laser beam is a function of the current flowing across the active junction. The spot size of laser beam becomes enlarge due to attenuation as it passes through the air around its environment. In this paper the available 3mW (milliwatt) laser diode is used in the laser beam transmitter circuit. It is class II type, 3 mW output power is in wavelength between 630 nm and 680 nm of visible region [5].

2.2 Characteristics of Laser Pointer

Laser pointers are manufactured in a variety of "colors" (wavelengths) ranging from green (532nm) to red (670nm). Our eyes respond to green light (~ 555nm). A green laser pointer appears brighter than a same power red laser pointer. The most common and inexpensive laser pointer wavelength is red(670nm). Class I lasers are those lasers which generate such a low powered beam that it cannot cause damage to anyone or anything. People can look directly into a Class I laser all day without any damage to their eyes.Class II lasers are those which have a visible beam with a power rating not exceeding 1mW. Class III lasers produce enough power to be dangerous to an eye. When a laser is directed into an eye the light is focused to a small spot on the retina and cause damage to the retina. These lasers are labeled with a red and white "DANGER" sign warning people to avoid getting the direct beam into the eye.Class IV lasers will produce serious eye damage even from scattered reflections of the beam. These lasers will also produce serious skin burns if a person is exposed to the beam [5].

3. DESIGN AND IMPLEMENTATION OF THE SYSTEM

3.1 Laser Guided Audio Transmitter

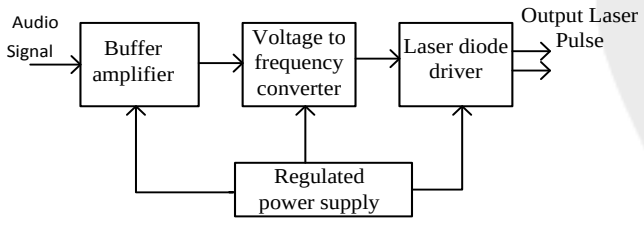


Figure 1. The basic block diagram of transmitter

The basic block diagram described in Figure 1 fundamentally composes of three parts: a buffer amplifier circuitry , a pulse position modulator , and the laser beam driver circuitry. The analog audio signal from the output of headphone is inputted to the buffer amplifier unit. The voltage to frequency circuit provides the frequency changes by modulating input voltages from the amplifier circuit. It can also be called the pulse position modulator. The last part is the current amplifier unit which amplifies the current by the action of transistor functions. The detailed circuit of transmitter section and it's part are also expressed in forwarding section.

3.2 Design of Audio Receiver System

Figure 2 is the complete block diagram of the receiver. The basic block diagram is a combination of the circuit of laser light detection by phototransistor, signal amplifier circuit , phase-locked loop design, low-pass filter circuit and the power amplifier circuit. The detailed functions and basic operations are expressed and also the necessity figures are also described in this section. The whole receiver unit is supplied by the constant voltage regulator supply voltage.

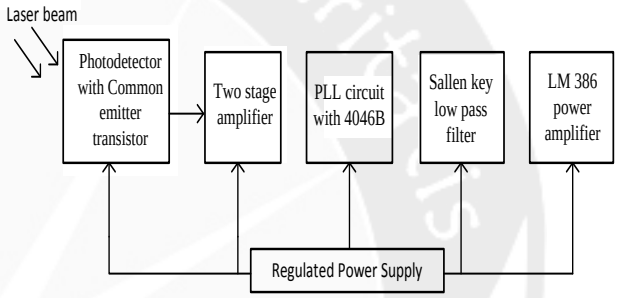


Figure 2. The block diagram of audio receiver

In the laser receiver circuitry, a phototransistor is used as the light detector and this provides good sensitivity at the higher frequency and in this case laser beam as the carrier frequency. The pulses of laser cause the collector-emitter resistance of phototransistor to fall slightly, and this generates the negative pulses at the collector of the transistor. The output from the phototransistor is low only a few millivolts or less. A stage of amplifier is essential to boost the signal. A high gain common emitter transistor serves as the initial state of the signal amplify stage.

The two-stage amplifying circuit is designed with LA 3161 and LM 741 op-amps for the signal amplifying stage. The task of an audio amplifier is to take a small signal and make it bigger without making any other changes in it. LA3161 has built in regulated voltage circuit so it is used to amplify the audio signal as a preamplifier. In the amplifying circuit, two op-amps are connected in series. The two amplifiers are functioned as inverting amplifiers The output of amplifying state is then connected to the input of phase-locked loop circuitry constructed by using 4046B PLL chip.

3.2.1 Light Sensor Used in Photodetector Circuit

The detection of optical radiation is usually accomplished by converting optical energy into an electrical signal. Optical detectors include photon detectors, in which one photon of light energy releases one electron that is detected in the electronic circuitry, and thermal detectors, in which the optical energy is converted into heat, which then generates an electrical signal. Often the detection of optical energy must be performed in the presence of noise sources, which interfere with the detection

process. The detector circuitry usually employs a bias voltage and a load resistor in series with the detector. The incident light changes the characteristics of the detector and causes the current flowing in the circuit to change. The output signal is the change in voltage drop across the load resistor.

3.2.2 Phototransistor

In this paper a phototransistor is used as a light detector. It is a transistor whose collector and emitter currents are directly related to the light incident on the base region of the transistor. Although any transistor will respond to light , the phototransistor has some features that make it more sensitive at a certain wavelength of light. When the base is illuminated with the correct wavelength of light, electron-hole pairs are formed in the base which creates a base current . Phototransistors are solid state light detectors that possess internal gain.The speed of response of a phototransistor is dominated almost totally by the capacitance of the collector-base junction and the value of the load resistance. A phototransistor takes a certain amount of time to respond to sudden changes in light intensity.

Table 1. Phototransistor voltage and current with variable DC voltages

Supply Voltage(V)	Voltage Drop(V)	Current(mA) with day light	Current(mA) with Laser light
2	1.69	1.25	0.25
3	2.73	1.5	0.25
4	3.7	3.7	0.46
5	4.8	4.8	0.56
6	5.76	5	1.2
7	6.15	5.2	2.3
9	10.7	5.7	2.6

3.3 Two-stage Amplifier

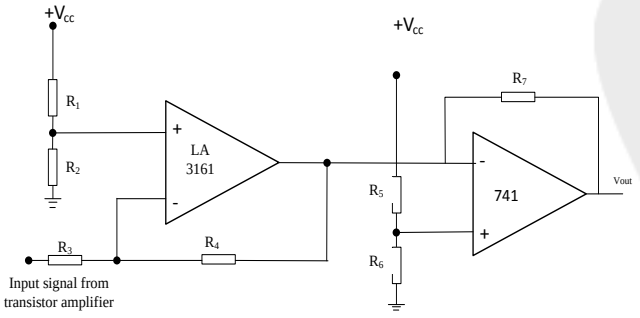


Figure 3. Two-stage amplifier

The two-stage amplifying circuit design with LA 3161 op-amp and LM 741 op-amp is shown in Figure 3. The task of an audio amplifier is to take a small signal and make it bigger without making any other changes in it. In the amplifying circuit, two op-amps are connected in series. The two amplifiers are functioned as inverting amplifiers. The input signal is applied through a series input resistor R to the inverting inputs.

The output is feedback through R_f. Negative feedback stabilizes against almost any type of disturbance. The gain can reduce to feedback a negative signal from the output which cancels parts of the input because negative feedback helps to overcome distortion and non-linearity. It also flattens frequency response.

The circuit properties are dependent upon the external feedback network and thus easily controlled by external circuit elements. The output voltage can never exceed the power supply voltage [2].

3.3.1 Function of 4046 B PLL in Receiver Circuit

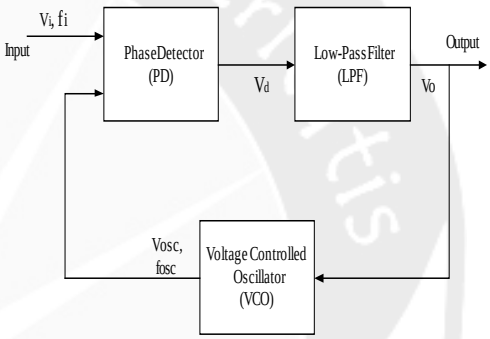


Figure 4. Basic block diagram of phase-locked loop(PLL)

Figure 4 shows the basic block diagram of phase locked loop. The input signal from the output of amplifier circuitry is applied to one of the phase comparator inputs of phase-locked loop circuitry pin (14). The 4046B can be used as PLL and not just a voltage-controlled oscillator (VCO). The center frequency is set approximately 100 kHz. The output from VCO is connected to phase comparator. The resistor and the capacitor pair formed a low pass filter between phase comparator's output and the input of VCO. The audio output signal is obtained from low-pass filter via an integral source follower stage. The voltage controlled oscillator input requires one external capacitor and one or two resistors determine the frequency range of the VCO.

The frequency capture range is defined as the frequency range of input signals on which the PLL will lock if it was initially out of lock. The frequency lock range is defined as the frequency range of input signals on which the loop will stay locked if it was initially in lock. The capture range is smaller or equal to the lock range.

3.4 Design of Sallen- key Low Pass Filter

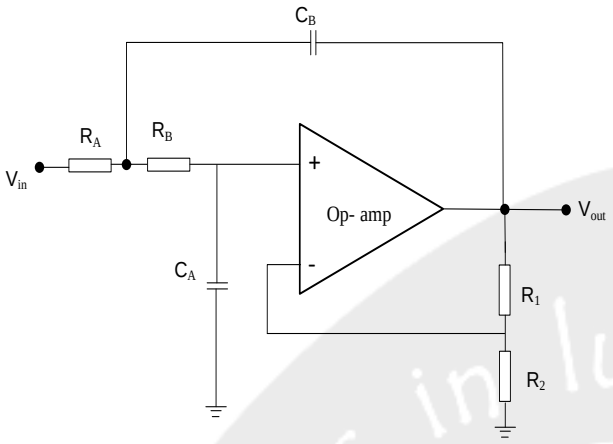


Figure 5. Sallen-key low pass filter

The Sallen-key filter is a popular filter due to it's versatility and ease of design. In this filter, there are two low pass RC circuits that provide a roll-off -40 dB/decade above the critical frequency. One RC circuit consists of R_A and C_A, and the second consists of R_B and C_B.

The critical frequency for 2nd order Sallen-Key filter is $f_c = (2\sqrt{R_A R_B C_A C_B})^{-1}$.The op-amp in the second–order Sallen-Key filter acts as a non-inverting amplifier with negative feedback provided by the R_1 / R_2 circuit. The damping factor is set by the values of R_1 and R_2, thus making the filter response. The R_1 /R_2 ratio must be 0.586 to produce the damping factor of 1.414 required for a 2nd order Butterworth response [3].

In Figure 3.5, non-inverting Sallen-Key is designed so that the input signal is not inverted. The low pass filter characteristic is shown in Figure 6. The gain option is implemented with R_A and R_B. The two resistors, R1 and R2 ,and the two capacitors C_A and C_B are connected to the op-amp's non-inverting input.

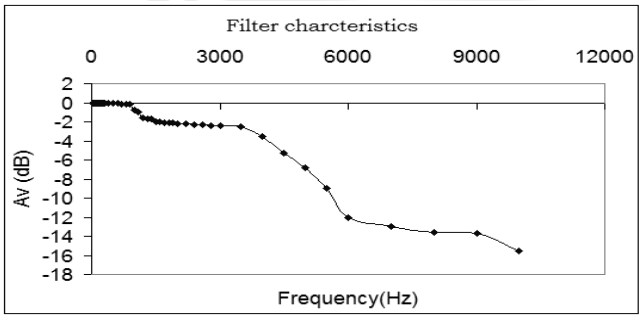


Figure 6. The curve for low pass filter characteristics

3.5 Power Amplifier LM 386 Op-Amp Circuit

The power amplifier supplies sufficient power to the headphone or speaker to enable the speech coil to move in and out by the required amount. A power amplifier should not be confused with a voltage amplifier which supply currents of up to several amperes and enable the speaker cone to move at frequencies of between 10 Hz and 20 kHz. Figure 3.7 shows the circuit diagram of LM 386 power amplifier circuitry as the part of the laser beam guided audio receiving system.The signal from the low-pass filter is fed to this power amplifier circuitry which provides an excessive output signal. Additional external components can be placed in parallel with the internal feedback resistors to tailor the gain and frequency response for individual applications. This amplifier will operate over a wide-range of voltage between 4V-12V [2].

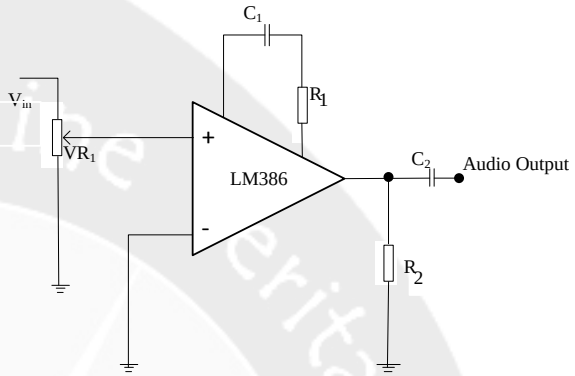


Figure 7. Power amplifier circuit

4. INTERFACING WITH ANALOG WORLD

4.1 Analog to Digital Conversion

Analog to digital conversion is an electronic process in which a continuously variable signal is changed, without altering its essential content into a multi-level signal.The input to ADC consists of a voltage that varies among a theoretically infinite number of values. The output of ADC has defined levels or states. The simplest digital signals have only two states, and are called binary. All whole numbers can be represented in binary form as string of ones and zeros. Digital signals propagate more efficiently than analog signals because digital impulses, which are well-defined orderly, are easier for electronic circuits to distinguish from noise, which is chaotic. This is the advantage of digital modes in communications. This paper aims to create hardware/software combination to convert audio signals into its digital counterpart [4]..

This is to be achieved by creating ADC hardware that is capable of converting at least twice the highest frequency of audio waves. To attain control of ADC device, the hardware circuit is connected via the parallel port, which is being controlled by the program software. In this thesis, the National Semiconductor ADC 0804 Analog-to-Digital Converter chip is used. The ADC 0804 uses the method of successive approximation to generate a digital code that is proportional to the applied analog input. The ADC requires a free running clock in a specific frequency range to convert drive the conversion circuitry.

4.2 Interfacing with PC's Parallel Port

For interfacing process the analog signal of the system must be converted to digital form. This action can be performed by analog to digital conversion process. The interfacing circuit is an essential part for the connection of real world and the computer. The purpose of the interface circuit is to convert the analog signal to

binary signal thatis suitable for the computer. The binary output from analogto digital converter will enter a specialized interfaced, a port. The program which is run by computer, will periodically examine the data input at this portand processing it**.** The parallel port of a PC has 8 data outputs and 9 handshake lines- 5 input lines and 4 output lines. This makes it easy to interface a wide range of user add-ons to a PC's port [6].

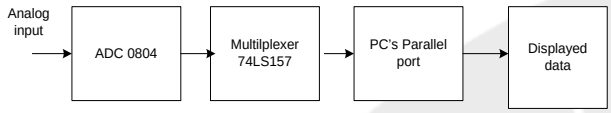


Figure 8. The basic block diagram of analog interfacing

For a given known input voltage, the software converts the corresponding binary numbers from ADC to ordinary decimal and it will be displayed on the screen. Input voltage can be varied DC voltage range of 0 to + 5V. Once ADC 0804 makes a digital conversion of the analog voltage on it's pins, the multiplexer 74LS157 is used to read a nibble data at a time, and then it switches to the other nibble and reads it. The software can be used to read the nibble from the Status port. After receiving two nibbles, into a byte raw value and display it on screen. The Status port is read and the most significant bit, corresponding to the Busy bit pin) is inverted using the exclusive-or function [1].

To make above hardware work, the multiplexer must be initialized first to switch either inputs low nibble or high nibble of 8 inputs from ADC. The least significant is read first and put A/B to low. The strobe is hardware inverted and then set Bit 0 of the control port to get a low on Pin 1 in parallel port . To start the conversion process WR is taken from low to high. Once the low nibble is selected, least significant bit can be read from the Status port. The Busy bit is hardware inverted and then XORed with 0x80 to toggle it back. The most significant nibble of the the results are only interested, and thus AND the results with 0xF0 to mask off the least significant nibble, and shift the nibble just read to the position of least significant nibble of variable. To get the most significant nibble, the multiplexer is needed to be switched to select input B and put the two nibble together to form one byte of digital value. This digital value can be returned to graphical interface to display it.[1]

Visual Basic program provides pre-built graphic user interface components and easily manipulate layout and component properties. VB also lets run the application while building it. The dialog box provides user to set sampling time. It also gives the option to choose running mode either just sampling the number of samples that user input or infinite sampling mode. The display can be stopped any time by pressing the Stop button.

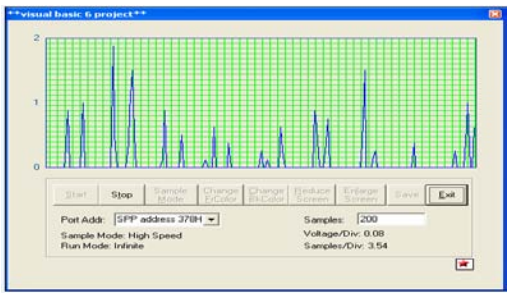


Figure 9. The graphical user interface of the system

5. CONCLUSION

Communication occurs when information is transmitted between an information source and the user of that information. There must be a transmission medium or channel between the source and the receptor . The laser beam is used as a communication medium. The range of the laser-based communication system may is fundamentally depend on the power the laser beam used in the transmitter and the sensitivity and the spectral response characteristics of the visible light detector. In this paper the audio signal is used as an information source and laser beam is applied as a high frequency carrier mediumwhere a carrier is a wave having at least one characteristic that may be varied from a known reference value by modulation. The intensity of the laser beam is being modulated to carry the information source to the other point of the receptor is placed. The receiving audio signal is needed to be converted in the form of digital so that it can be interfaced and the digitized signal isvisualized on the PC's screen provided by VB programs. The ADC 0804 hardware circuitry is built and connected to the PC's parallel port to process the visualization of data. The data visualization is monitored on the PC so as to detect any interference between the transmitter and receiver circuit.. The system can be useful in optical communication as well as data visualization applications. The performance of this system will be better if the high output power of laser diode can be used and the light detector of having a wide sensing area and the fiber optic cable is used as the communication medium.

6. ACKNOWLEDGMENT

I wish to express my deepest gratitude to Dr. Pike Tin, (Retired) Rector of University of Computer Studies Yangon for his continuous interest in this study.Special Thanks to Daw New Ni , Professor and Head of Hardware Department of University of Computer Studies Yangon for her kind permission to use the required laboratory instruments in this research and gave valuable suggestions at the seminars and made it possible for me to complete the whole thesis.Finally, I also wish to thank my parents who encourage and gave strength during my thesis and all friends and persons who directly or indirectly contributed towards the success of this thesis.

7. REFERENCES

[1] Parallel Port Shark Project Comunicazione TRA Personal Computer, Andrea Battistotti, Armando Leggio ,Copyright 1997-2001 Craig Peacock - 19th August 2001.

[2] Philip semiconductors Product Specification http://www.datasheetarchive.com

[3] Sallen-Key Low_Pass Filter http://www.ecircuitcenter.com/Circuits

[4] Analog to Digital Conversion http://www.ltl13.exp.sis.pitt.edu/ADDA paper.htm

[5] Laser Application http://www.EncyclopediaBritannica6.htm

[6] PC Interfacing http://delabs.tripod.com/data2.html

[7] Laser Communication http://www.richland2.k12.sc.us

Development of Steganography Software with Least Significant Bit and Substitution Monoalphabetic Cipher Methods for Security of Message Through Image

Iswar Kumbara
Computer Science Faculty
Sriwijaya University, +62-711-379249
Informatics Department, Palembang, Indonesia
iswarkumbara@gmail.com

Erwin
Computer Science Faculty
Sriwijaya University, +62-711-379249
Informatics Department, Palembang, Indonesia
erwin@unsri.ac.id

ABSTRACT
Steganography is a science and art which explains about how to hide secret information in to another media that make people does not realize the existing of this message. But steganography has a significant weakness, that is the operation has been widely known. Therefore the writer tried to combine two methods; The first method was Least Significant Bit for Steganography and the second was Substitution Monoalphabetic Cipher for Cryptographic. The combination of the two methods was carried out into media picture 24 bit due to this media was not used up the resources memory. So that, it is very possible to send image media through email. By applying the two methods, it is hope that the message's security is safe from the disturbance of other people who does not have authority to read the message.

Keywords
Steganography, Least Significant Bit, Monoalphabetic Cipher Substitution, Image.

1. INTRODUCTION
Message is something that is very secret and should be known only by the sender and recipient, therefore the sender does not want if the contents of their message are known by other people who does not have any authority to read the message. By seeing this condition, the writer was trying to build a message security through image application. Why using this media? Because image media is commonly used when someone wants to send picture through email, and its size is not too big, so it is easy to send. then, another question appear : why should a message be hidden if media mail has been very secure?

An idea was found when the writer was working in one of leading state-owned enterprise in Indonesia, where its network security still have problem, as reading someone's email is not so difficult. However, the discussion about reading the other person's email messages is out of the writer's boundary.

The commonly used method for securing message is cryptography method, which is used for security system with encrypting data inside. But, the disadvantage of this method is that everyone can see that the message has a confidential and it will raise the desire for someone (so called cryptanalyst) to break the confidentiality in the message. For this reason, a developed method from Cryptography, Steganography, is emerged. But, it also has drawback in terms of data security; the lack of this method is the quality of the data as a container for message will be disturbed (Munir, 2004). These deficiencies can arise the desire of so called-steganalist to dismantle the message, after they felt sure that the file has a hidden message

To minimize the weakness, the writer tried to combine the method of cryptography and steganography for a more complex security message. The containers used to hide the message, is image media as described at the beginning of the introduction.

This paper is based on previous study ever with title "Making Application Steganography (Embedding Message in Picture)" by Kurnia Windyartinigsih in 2007.

2. FUNDAMENTAL THEORY
In this software development, cryptography process was done and followed by steganography process. Figure 1 shows the process details.

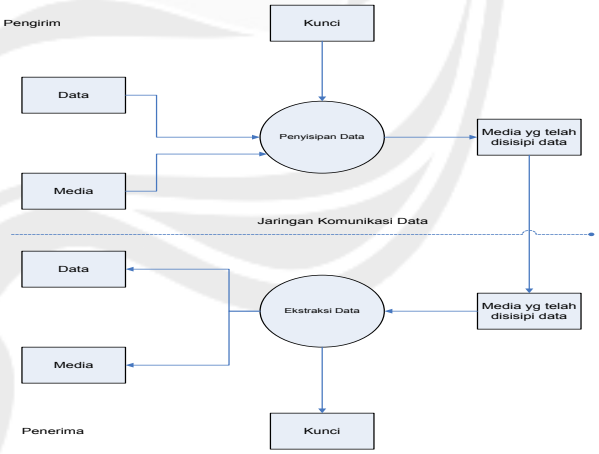
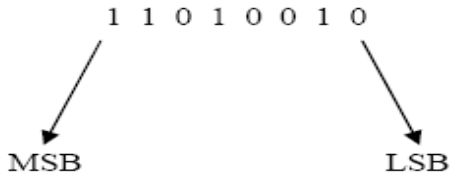


Figure 1. Inserting message system diagram

3. STEGANOGRAPHY
3.1 LSB Method
There are two steps in steganography system. They are hidden process and recovery data from the file repository. Data hiding is done by replacing the bits of data in the segment images with multiple bits of data secret. In arrangement of bits in a byte (1 byte=8 bits), there is the Most Significant Bit (MSB), and the

Least Significant Bit (LSB). The least significant bits will be replaced with a bit of the secret message.



Bits suitable for this change is LSB, because that change is just change one byte value higher or one lower than the previous value. For example, if the byte supposed to show red color, the change one bit LSB will not change the red color in a significant way. Afterall, the bare human's eyes can not identify such small changes.

Example of the image data segment before change:

0 0 1 1 0 0 1 **1** 1 0 1 0 0 0 1 **0**

1 1 1 0 0 0 1 **0** 0 1 1 0 1 1 1 **1**

Image data segment after '0 1 1 1' are hidden:

0 0 1 1 0 0 1 **0** 1 0 1 0 0 0 1 **1**

1 1 1 0 0 0 1 **1** 0 1 1 0 1 1 1 **1**

So the data that are intended to be inserted, by replacing the last bit.

3.2 The Data Size in Image

The size of hidden data will depend on the size of the image containing it. In 24-bit image which has size 256*256 pixel, there are 65536 pixel. Each pixel's size is 3 bytes (Red, Green, Blue components), means a total of 65,536*3=196,608 bytes. Because each byte can only hide one bit in its LSB, then the size of the data to be hidden in the image maximum is 196608/8= 24576 byte. The greater data hidden in one image, the greater possibility of damaged data in the image manipulation contain.

4. MONOALPHABETIC CIPHER SUBSTITUTION METHOD

Monoalphabetic Cipher Subtitution Method was derived from the caesar method used by Roman emperor Julius Caesar, but because it is easy in dismantling the lock on the method, monoalpabetic cipher substitution method was developed. The principle of caesar method is to shift each letter that include as many as three, or the amount of shift is k=3. So if we write A alphabet so the output is D alphabet. Monoalphabetic cipher substitution of replacing the letter with another letter or a specific symbol, depending on the number of letters and symbols available.

For Example Letter Becoming Key :

Letter	Key
a	!
b	@
c	#
d	$
e	%
f	^
g	&
h	*
i	(
j	?
k	>

5. INSERTION KEY MESSAGES

The process of inserting the key message are:

1. subtitute the original message into a key message.

```
if(message[x]= 'a')
        message[x]='?';
 else if(          )
```

2. The subtitute message was inserted to an Image which is already loaded

```
BinaryMessage=BibaryMessage+binary
(AsciiToInt(message[x]));
```

3. Save Image

6. READING MESSAGE IN IMAGE

The process is the anthithesis of the image insertion process, namely:

1. Image contain message is loaded, continue with reading the message on that image

```
message=message+chartoascii
(decimal(BinaryMessage.SubString(n,8)));
```

2. The message resulted from the image reading is subtituted by key messages

```
if(message[x]=='?')
message[x]='a';
else if (          )
```

3. The message included in the process insertion of the key messages can be read, according to its original form.

7. SOFTWARE ANALYSIS

Software interface consist of Login Form, Change Password Form, Steganography Encryption Form, and Steganography Decryption Form.

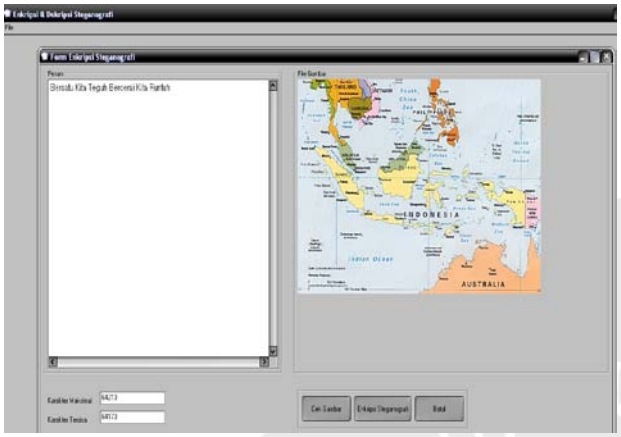


Figure 2. Steganography encryption form

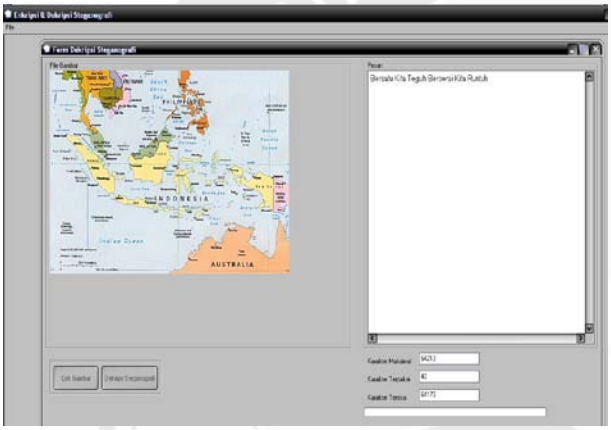


Figure 3. Steganography decryption form

8. IMAGE TESTING

The Message inserted in Figure 4 was "Bersatu Kita Teguh Bercerai Kita Runtuh." after the message was inserted, the image did not visibly changing because bit that were replaced not too significant in the image quality. See Figure 5

Figure 4. Image not inserted message

Figure 5. Image after inserted message

In the test above, the usage of cryptography would not seen because the message that already crypt has been opened at the time of decrypting the message.

Figure 5 shows that the message on the image was opened using LSB method only.

If the image was opened by only using the LSB method, the message obtained would be

s=8#?6@Wl|6?WV=>@<Ws=8[=8?|W l|6?WC@96@<*

and not "Bersatu Kita Teguh Bercerai Kita Runtuh"

10. TESTING IMAGE WITH HISTOGRAM

In figure 4 and figure 5 the writer compare between the two images, which is seen by the naked eye. Surely the human eye can not notice a difference between the second image, which is why the author uses a histogram to show whether there is a significant difference between the image that has been inserted with the message that the message has not been inserted. See Figure 6

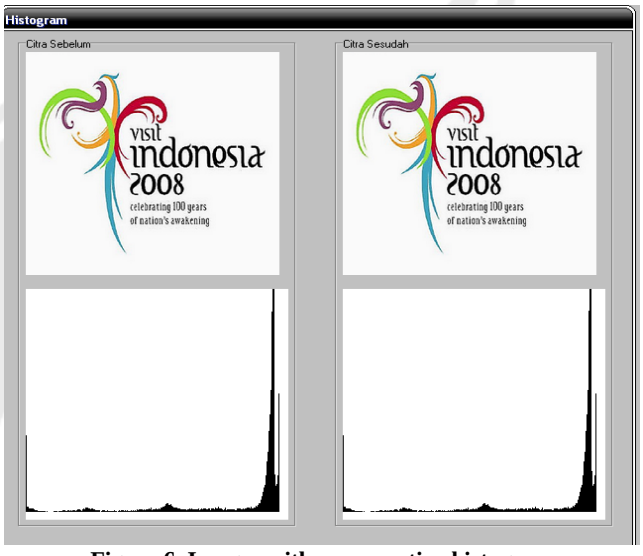


Figure 6. Images with comparative histogram

Of Figure 6 shows that by using the histogram differences between the two images are not seen to be significantly different.

11. CONCLUSION

Based on the results of the analysis, a software tool to inserting and extracting message into image has been successfully developed, and before the message is inserted into the image, the message will get cryptographic process first. Although someone has managed to uncover a secret message that has been inserted, they will get only the messages that can not be read.

In LSB method, only the last bits are going to be changed, because using the other bits has high probability to reduce the image quality.

Both methods still can be developed with the addition of other methods, either using a random function on the pixels or use the XOR function, before the bit is inserted in the message. The more security is applied, the safer the message will be.

12. REFERENCES

[1] J. Cox, Ingemar, et all. 2007. Digital Watermarking and Steganography. USA :Morgan Kaufmann

[2] Masaleno,Andino.2006.Pengantar Steganografi.Online : www.ilmukomputer.com. [18 Maret 2009]

[3] Munir, Rinaldi. 2004. Pengolahan Citra Digital.Bandung : Informatika Bandung.

[4] Kurniawan, Yusuf. 2004. Kriptografi Keamanan Internet dan Jaringan Komunikasi. Bandung : Informatika Bandung.

Feasibility Analysis of Zigbee Protocol in Wireless Body Area Network

Vera Suryani
Informatics Engineering Dept.
Institut Teknologi Telkom
Jl. Telkomunikasi no 1
Bandung, West Java, Indonesia
+62227564108 ext 2107
vra@ittelkom.ac.id

Achmad Rizal
BioSPIN RG
Institut Teknologi Telkom
Jl. Telkomunikasi no 1
Bandung, West Java, Indonesia
+62227564108 ext 2014
arl@ittelkom.ac.id

ABSTRACT
In this paper, we describe feasibility analysis of Zigbee protocol in Wireless Body Area Network (WBAN). We use NS2 to simulate data from oscilloscope as model for physiological signal from sensor to WBAN's server. Four sensors were mounted in patient body and send the data simultaneously to the server using Zigbee protocol. Simulation result shows that packet loss, throughput and delay of the data transmission using Zigbee protocol is feasible to use in WBAN's data transmission.

Keywords
Zigbee protocol, WBAN, Telemedicine,

1. INTRODUCTION
Telemedicine system can be interpreted as a technique / method to perform medical procedures in the overall distance. In the implementation, telemedicine telecommunications infrastructure needed to connect the parties involved in the telemedicine (doctors, patients, pharmacists and others). The simplest form of telemedicine is the consultation between patients with a doctor using telecommunications services. In this mode, the doctor can only determine the patient's diagnosis of the complaint could not check the condition of the patient directly. To ensure valid diagnosis, where doctors can examine the patient's physiological signals, needed a mechanism to acquire the patient's physiological signals and send them via telecommunications networks. Telemedicine system like this would require a series of sensors to acquire physiological signals from the patient and sent to doctor's

PC. If there are several sensors mounted on the patient's body and simultaneously send a physiological signal data, sensor networks will be required. Series of sensors mounted on the body which must be integrated in one network so you can easily set in the data transmission.

A body-centric network, so-called BAN-Body Area Network, can be formed by integrating these devices on a human body (or its proximity). If BAN is implemented wirelessly, so we call it WBAN-Wireless Body Area Network. WBAN, with sensors consuming extremely low power, is used to monitor patients in critical conditions inside hospital. Outside the hospital, the network can transmit patients' vital signs to their physicians over the Internet (or private networks) in real time. In WBAN implementation, some parameters must be considered such as chosen platform. After we chose the platform, we have to measure the other network parameters such as delay time, packet loss, throughput and so on.

2. WBAN USING ZIGBEE
Wireless body area sensor networks (WBANs) are well suited to increase telepresence, as they can provide specific information about an individual's behavior without using complex laboratory equipment and without interfering with the person's natural behavior [11]. WBANs are generally built around several sensing devices wirelessly linked together using narrow-band radio communication [12]. Recent developments in the field of wireless networks have generated many new commercial wireless communication platforms based on different protocols and technologies (Wi-Fi, WiMax, Bluetooth, Zigbee, UMTS, UWB) [13]. These technologies offer a wide range of characteristics in terms of speed, transmission range, power requirements, connectivity, and cost. The choice of wireless network architecture for a WBAN application is context and sensor dependent.

The use of a WBAN system in telemedicine context calls for a small, reliable, low-power platform capable of seamlessly integrating several modules. The Zigbee technology was designed for this type of application. The IEEE 802.15.4 physical radio standard operates on the 2.4-GHz unlicensed band over 16 channels, and the network layer supports topologies such as star, tree, and mesh. Depending on the power output and environmental characteristics, transmission distances range from 10–100 m [14].

ZigBee's main advantage is its ability to be configured in so-called mesh networks with wireless nodes that are capable of multi-year battery lives. In a mesh topology, each node is in direct communication with its immediate neighbor; if a node fails, messages are automatically rerouted a sort of miniature Internet. ZigBee also supports more efficient star topologies, in which central access points talk to the nodes.

ZigBee is actually the network protocol, security, and application layers for one type of network that can run on radios conforming to the 802.15 standard of the Institute of Electrical and Electronics Engineers Inc. (IEEE), an umbrella that also covers Bluetooth and other types of wireless personal area networks (WPANs). The physical layers for ZigBee transmitters are described in IEEE 802.15.4 and were approved in 2006.

Table 1. Possible BAN/WBAN platformsWireless technologies and possible BAN/WBAN platforms [1]

Technology	Transfer Rate	Range	BAN/WBAN
WiFi	11 – 54 Mb/s	30 – 50 m	PDAs
WiMax	45 – 70 Mb/s	100 m – 50 km	Portable computers
Bluetooth	57 kb/s – 3 Mb/s	100 m	iMotes
Zigbee	20 -250 Mb/s	100 m	MiCaz, Telos
UMTS	50 kb/s – 2 Mb/s	5 – 100 km	Mobihealth
UWB	54 kb/s – 48 Mb/s	1 – 10 km	Magnet

The Zigbee protocol was designed to be optimal for the control and sensor application space. It is less complex than Bluetooth, has superior power management (2 AA battery can have ZigBee module last over years) , supports many more nodes per network, has lower latency, lets devices join the network more quickly, and wakes up in milliseconds instead of seconds. These of advantages of Zigbee, make it become the right platform for WBAN's application.

3. MODELLING AND SIMULATION

The simulation of WBAN in WSN network using IEEE 802.15.4 was conducted by *Network Simulator 2* (NS2). The inputs for the simulation are describe below.

3.1 Topology

Mesh topology is chosen in the simulation because each sensor is directly connected to the PAN coordinator. That's why no routing protocol used in scenario. Topology is selected with 4 sensor nodes and 1 PAN coordinator. All of the sensor nodes attach on 1 patient but located in different coordinates. Topology used in the simulation can be seen in Figure 1.

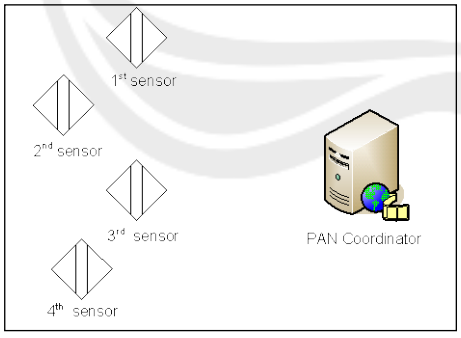


Figure 1. Network topology used in scenario

3.2 Scenario

The input parameters for the simulation are presented in Table 2.

Table 2. Input Parameter for Simulation

Parameter	Spesification
Distance between nodes	± 25 meter
Packet size	30 bytes
Packet rate	5 Kbps
Duration	900 s
Propagation	Two Ray Ground
Routing protocol	None

These inputs are used in the scenario for measuring: packet loss, throughput and delay between sensor nodes to PAN coordinator.

Delay = propagation delay + transmission delay + queuing delay

$$\text{Packet loss} = \frac{\text{packets that fail to reach their destination}}{\text{all packets sent from source}} \times 100\%$$

$$\text{Throughput} = \frac{\text{successful packets reaching their destination}}{\text{all packets sent from source}} \times 100\%$$

4. SIMULATION AND ANALYSIS

4.1 Delay

Each sensors yield different delay, as shown in table 3.

Table 3. Delay from Sensor to server

Source	Destination	Delay (ms)
1	0	0.352034
2	0	0.352031
3	0	0.352033
4	0	0.352032

The average delay for all sources is calculated as 0.3520325 ms. It means that each node only needs to transmit directly to destination because the routing protocol is disabled. The delay will increase if we use routing protocol such as AODV for finding path from source to destination.

4.2 Packet loss

Each nodes has packet loss when sending the packets, as shown in table 4.

Table 4. Packetloss from all nodes

Source	Packetloss (bytes)	Percentage (%)
0	101640	1.65595445
1	87150	1.4198783
2	85080	1.38615313
3	93690	1.52643026
4	96900	1.57872871

From the tracing file during simulation, it shown that drop packets caused by failed connection establishment in TCP connection. This happened because of low signals from the source to destination. None of the packets contain data is drop during the transmission.

4.3 Throughput

The received packets for each node presented in table 5 :

Table 5. Throughput from all nodes

Source	Throughput (bytes)	Percentage (%)
0	6036210	98.34404555
1	6050700	98.5801217
2	6052770	98.61384687
3	6044160	98.47356974
4	6040950	98.42127129

From table 5, we can infer that throughput from each sensors and PAN coordinator yield in for about 98% from overall packets transmitted. And the packets received are data, which means that Zigbee protocol is reliable enough to be used as network technology for WBAN.

5. SUMMARY

The average delay for all sources is calculated as 0.3520325 ms

Drop packets caused by failed connection establishment in TCP connection, furthermore it's also caused by low signals from the source to destination.

Throughput from each sensors and PAN coordinator yield in for about 98% from overall packets transmitted, means that Zigbee protocol is reliable enough to be used as network technology for WBAN.

6. REFERENCES

[1] Etonnet,2007, Zigbee Advantages, Etonnet Inc. DOI=http://www.etonnet.com/zigbeeadvantage.aspx

[2] Otto, Chris., Jovanov., Emil., 2006. An Implementation of the WBAN Health Monitoring Protocol for ZigBee Compliant TinyOS Messaging, University of Alabama in Huntsville, DOI= *www.ece.uah.edu/~jovanov/projects/ WBAN_HM_Protocol.pdf*

[3] UC Berkeley, LBL, USC/ISI, and Xerox PARC : The ns Manual, March 2008.

Mobile TV with RTSP Streaming Protocol and Helix Mobile Producer

Yunianto Purnomo
Informatic Engineering Department
Krida Wacana Christian University
Jl. Tanjung Duren Raya no.4, Jakarta
+628161839260
yunianto@ukrida.ac.id

Andrew Jaya Efendy
Informatic Engineering Department
Jl. Wijaya Kusuma Raya no.59
Taman Yasmin, Bogor
+628381010229
andrew@maxindo.net

ABSTRACT

The precense of technology provide people to grab information easy and comfortable. A lot of media are available to be used in order to get information very fast such as mobile phone. Nowadays mobile phone can be used to watch television too.

Television broadcast are transmitted via internet that does not require special infrastructure development. Online internet TV can be watch from mobile phone using the "Darwin Streaming Server" technology from the internet server as a broadcast TV service provider.

Mobile phone which can receive online TV broadcast installed "Real Player Mobile" software or other software that supports RTSP file format. RTSP Streaming protocol, H263 video codec, and the AMR codec are used by online internet TV application. By using digital technology, it is expected that TV viewers could watch TV program in perfect picture quality anywhere even in moving vehicle. It is also a simple way for television stations to broadcast a television broadcast without having to build many transmission towers.

Keywords

Mobile TV, RTSP Streaming Protocol, Helix Mobile Producer.

1. INTRODUCTION

Indonesian territory scratch from Sabang to Merauke almost two millions square kilometers and consist of 17508 islands. To develop analogue TV network in this huge area need time and much effort. That's why we should develope digital television network as a solution to the analogue television network.

Nowadays are available digital TV services that broadcast via satellite or cable TV, but it is relatively expensive to pay the monthly fee. On other had we have to develope infrastructure such as parabola dish and decoders to receive this service.

Therefore a new breakthrough by using internet TV. This network does not require special infrastructure because it was provided by internet infrastructure. So it is free of charge for watching television broadcast.

Sample of mobile TV is shown on figure 1 below:

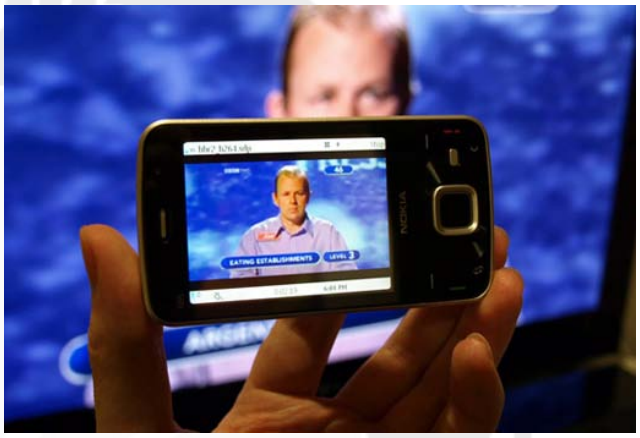


Figure 1. Mobile TV Sample

2. CURRENT TRENDS

Trends in television broadcasting is switching from analog to digital broadcast. Using media such as mobile phones, personal digital assistance (PDA), digital TV offers better quality.

Communication and Information Ministries in May 2009 lead a team conducted digital television broadcast trial for mobile internet TV receiver and will make a trial in fixed TV.

Digital broadcast for mobile TV trial was conducted by "Tren Mobile TV" consortium and the "Telkom-Telkomsel-Indonusa" consortium since August 3rd, 2009 by using OMABICASS system. The transmitter attached on Menara Kebon Sirih building with coverage Central Jakarta using 24 UHF channel.

10 TV programs broadcast by "Tren Mobile TV", those are: TVRI, RCTI, TPI, Global TV, MNC News, CNN, Al Jazeera, Bloomberg, MNC Music, and MNC Entertainment. Communication and Information Ministries appointed community representative to distribute 50 Nokia mobile phone type N77 for trial in receiving DVB-H broadcasts.

"Tren Mobile TV" has signed a memorandum of understanding with BPPT to measure the strength and quality of this broadcast signals. This was made to accelerate development of Mobile TV in Indonesia with new standard in providing Mobile TV.

Public received analog broadcast which quality depend on the strength of broadcast frequencies. Digital TV broadcast will provide sharp, perfect, and stable image even in moving vehicle and in bad wheater. This quality was enabled by digital TV

broadcast technology that convergence between image (video), data (internet) and voice.

Now digital TV broadcast views by some people in Jakarta and need improvement to deploy digital TV broadcast throughout Indonesia.

Figure 2. Streaming Video & Audio Work

3. THE APPLICATION

Development of the internet TV is not using UHF frequencies as transmission media, but using RTSP protocol over internet connections. This is very advantageous, because by using the Internet connection is not disrupted by the weather, and the interference signal at the receiver antenna. Another advantage of RTSP protocol is supported by almost all newest mobile phone, so it is no special phone is required to receive television broadcast.

The computer server of internet TV requires only 1 unit computer installed some support software, those are: Encoder Helix Mobile Producer, Darwin Streaming Server, and equipped with static IP internet connection. a server can only broadcast a television channel. To broadcast multiple channel have to use some computer server.

Server of internet TV works by capturing and broadcast a channel of television in real time synchronous. The simple steps are: a computer server with TV tuner capturing broadcast television (for example: TVRI as TV provider) through the receiver antenna with a certain frequency or directly via a cable if it is done at the TVRI station. TV tuner transmit the television broadcast to Helix Mobile Producer encoder then it trans coded (format changes) to a 3GP format file. Helix Mobile Producer encoder transmits the 3GP data packet video streaming to Darwin Streaming Server for broadcasting through an internet connection with the RTSP protocol. Mobile phones get internet TV broadcasts by connecting to the server for a 3GP data packet streaming via internet which is displayed on the video player phone screen using a software that supports streaming over RTSP protocol. Example of the software were installed in mobile phone to show it is Real Player Mobile. This software is pre-installed by vendor.

By using method mentioned, the broadcast of internet TV does not require a digital transmitter as practiced by Trends Mobile-Telkomsel and Telkom TV-Indonusa. Using the transmitter becomes limited broadcast range area and it requires an expensive cost for construction.

4. INFRASTRUCTURE

There are three main infrastructure for internet television broadcasts: the mobile phone network with a minimum of 50kbps bandwidth internet, television service providers and a mobile phone that supports XHTML programming languages.

4.1 Network

3G technology enables mobile network operators to provide service to users for a broader range of internet with the ability to display video calls. 3G technology consisting of three standards: Enhance Data rates for GSM Evolution (EDGE), Wideband-CDMA, and CDMA 2000. The weakness of this technology is relatively more expensive and lack of network coverage because this technology is still new.

EDGE is the minimum standard requirement to support the successful of mobile TV services. EDGE support data transmission speeds up to 384kbps.

HSDPA is the latest technology in mobile telecommunication systems, issued by the 3GPP Release 5. This is a 3.5-generation technology (3.5 G). This technology is an enhancement of WCDMA. Similarly with CDMA2000, it developed EV-DO which is designed for high speed data transfer.

HSDPA has a data packet based services in WCDMA downlink with data rate up to 14.4 mbps and 5 MHz bandwidth in WCDMA downlink. For streaming, data services are more widely used in the downlink than the uplink.

4.2 Server

In the computer server, there are applications that work for a broadcast television. It uses PHP and XHTML mobile as a programming language for display application, RTSP protocol for video transmission, Darwin Streaming Server is an application server to transmit the RTSP protocol, Helix Mobile Producer is the encoder application to retrieve and send the impression from the TV tuner to Darwin Streaming Server.

This application is designed using RTSP protocol that regulates the transmission of video and audio for this protocol suitable for video and audio is compressed from the original size to a smaller size. This protocol also provides a connection that has the status of servers and clients, this will simplify the client to pause or find a random position in the stream when playing back a video.

Streaming application servers used in this server machine is Darwin Streaming Server (DSS). DSS was selected because this application is freeware. The application supports RTSP streaming protocol and supports Live Stream, which is streamed directly from an existing multimedia device such as TV tuner, webcam, and others using encoder applications.

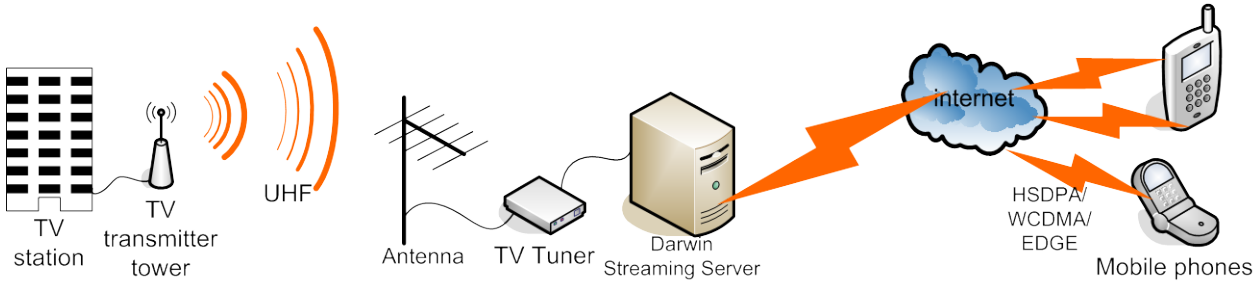


Figure 3. From TV station transmit by antenna to server

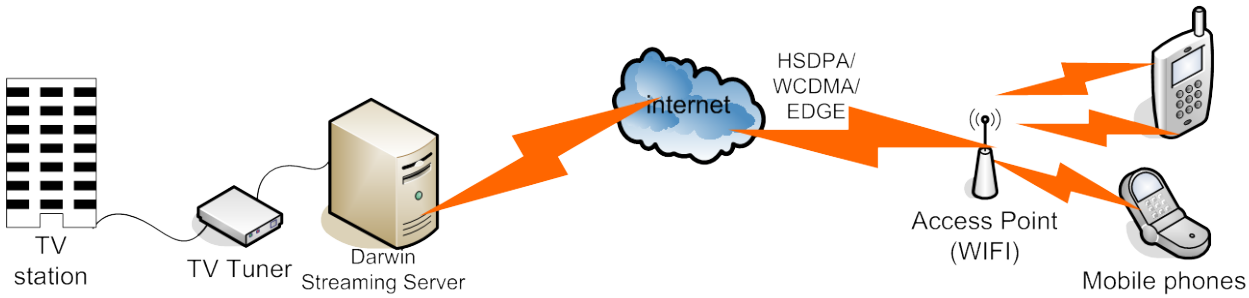


Figure 4. From TV station transmit directly to server

Encoder application used Helix Mobile Producer. Although the encoder application is a commercial with a fairly expensive price, the features are needed such as changes in video and audio codec to H.263 and AMR-NB which is a codec that can read most mobile devices, capture video and audio from a variety of multimedia devices, and transmit video and audio to users via streaming server application.

4.3 Mobile Phone

This device is used to receive internet television broadcasts / mobile TV. Viewers can watch TV anywhere via mobile phone the screen. The mobile phone is meant to capture internet television broadcasting / mobile TV is not mobile phone that has a TV feature that captures TV broadcasts via UHF or VHF signal.

Not all types of mobile phones can be used to watch this television broadcast. Mobile phone used to watch television broadcasts online / mobile TV must have features that can capture EDGE or HSDPA signals, which can be installed "Real Player Mobile" software or other software that supports RTSP file format, and support XHTML. One of the phones which has these features is a Nokia N77.

5. SCOPE

Every region in Indonesia has many analog TV channels. Each television stations broadcast a channel at a time generally. Only a few television stations broadcast more than a channel at a time, like TVRI.

Similarly, the design of this internet television / mobile TV: an internet television station generally uses a server device to produce a channel broadcasts. A server device can only broadcast a channel only. For an internet television broadcasting service provider / mobile TV intends to broadcast more than 1 channel, it must prepare more than a server device.

6. IMPLEMENTATION

Something necessary to be prepared for the internet television broadcasting for internet mobile TV:

6.1 Streaming Server Setting

It requires a streaming server to relay (forward) video or audio from an encoder, transforming into a video or audio input to a format that can be played on mobile devices. Streaming server for this purpose are Mobile Streaming Server (MSS).

Setting to be done:

- Set the IP and port by the MSS.
- Set the IP encoder that will be forwarded by the MSS.
- Set the relay MSS to determine the origin and destination streams of data packets.

After that the MSS is ready to receive data packets transmitted from the encoder.

6.2 Encoder Setting

The encoder is used to change the input from TV tuner into a video and audio formats so that it can be played by mobile devices. Encoder that can do this is the encoder that has the same codec with codec on the mobile device using the RTP protocol to transport data. This is called mobile encoder.

Setting to be done:

- Determining the video and audio source (file or device).
- Determine the codec used.
- Determine the bit rate of the output.
- Define the output result (file or streaming RTP).
- Defining the IP and port of the streaming server if the output is an RTP stream.

After setting is done, then the encoder is ready to transmit data packets streaming video and audio from one device to the mobile device.

6.3 Mobile Phone Setting

To watch television broadcasts, viewers using mobile phones whit specifications to receive EDGE or HSDPA signals, and a "Real Player Mobile" software or other software to play the video as a television broadcast.

Use the cellular provider that offers 3G or 3.5G (EDGE or HSDPA) services. Arrange the features of 3G or 3.5G in the phone accordance with the specifications of the cellular provider. Today, almost all cellular phone service provider is provide 3G or 3.5G services. The difference of them is in setting of their specification.

6.4 Other Setting

For Internet television broadcasting service provider / mobile TV needs to prepare the server device and also set up internet network connection with data upload speeds of 2mbps, or at least 50kbps. Server takes a very high speed to upload video data.

Figure 5. Menu on Cellular

7. CONCLUSIONS

This mobile TV / Internet television is still in experimental. Wishing it will come true a digital television broadcasts soon, so that it can be used by most people via mobile phone anywhere.

Digital television via mobile phone network will enable to view broadcast television with excellent image quality with that not affected by the weather.

It is also a simple way for television stations to broadcast a television broadcast without having to build many transmission towers.

Internet television / mobile TV can improve cellular phone performance in the future. Cellular phone which is generally used for voice communication, text communication (SMS), and data communications (internet), in the future will be used as a tool for watching television either.

8. ACKNOWLEDGMENTS

Our thanks to PT. Maxindo Mitra Solusi and Krida Wacana Christian University – Computer Laboratory for allowing us to make an experiment about mobile TV / internet television, using your server, internet network, etc.

Figure 6. Film on Cellular

9. REFERENCES

[1] Commer E. D, 2004, Computer Networks and Internets with Internet Application. 4th edition, Pearson, Prentice-Hall, New Jersey.

[2] Nathan J. Muller, 1998, Telecommunications Factbook, McGraw Hill, USA.

[3] Tanutama L, 1989, Pengantar Komunikasi Data. PT. Elexmedia Komputindo, Jakarta.

[4] Wahyono T, 2003, Prinsip Dasar dan Teknologi Komunikasi Data. Graha Ilmu, Yogyakarta.

[5] William C. Y. Lee, 2006, Wireless and cellular telecommunications 3rd edition, McGraw Hill, Singapore.

[6] http://en.wikipedia.org/wiki/Mobile_television, 7 feb 2010.

[7] http://id.wikipedia.org/Enhanced_Data_Rates_for_GSM_Evolution, 7 feb 2010

[8] http://id.wikipedia.org/GPRS, 7 feb 2010

[9] http://id.wikipedia.org/High-Speed_Downlink_Packet_Access, 7 feb 2010

[10] http://id.wikipedia.org/wiki/Handphone, 2 feb 2010

[11] http://id.wikipedia.org/wiki/Televisi_internet, 2 feb 2010

[12] http://indonesian.red5server.org/selected_news_500081, 3 feb 2010

[13] http://wartawarga.gunadarma.ac.id/2009/10/evaluasi-tools-kompresi-file-multimedia/, 3 feb 2010

[14] http://www.total.or.id/, 7 feb 2010

[15] http://www.total.or.id/info.php?kk=Video%20Streaming, 8 feb 2010

Quantitative Performance Mobile Ad-Hoc Network Using Optimized Link State Routing Protocol (OLSR) and Ad-hoc On-demand Distance Vector (AODV)

Andreas Handojo
Department of Informatics Engineering, Faculty of Industrial Technology, Petra Christian University
Siwalankerto 121-131, Surabaya 60236, Indonesia
+62-31-2983455
handojo@petra.ac.id

Justinus Andjarwirawan
Department of Informatics Engineering, Faculty of Industrial Technology, Petra Christian University
Siwalankerto 121-131, Surabaya 60236, Indonesia
+62-31-2983455
justinus@petra.ac.id

Hiem Hok
Department of Informatics Engineering, Faculty of Industrial Technology, Petra Christian University
Siwalankerto 121-131, Surabaya 60236, Indonesia
+62-31-2983455

ABSTRACT

Information transfer is one of a major issue in information technology development. This is because one of basic purpose of development of information technologies is intended to transfer information between the parties. One of the latest developments in information transfer is the routing called MANET (Mobile Ad-hoc Network) which is used as one standard routing on the wireless world. MANET itself is divided into two methods are proactive routing method, which is represented by the Optimized Link State Routing (OLSR) and reactive routing method, which is represented by the Ad-hoc On-demand Distance Vector (AODV). In this research, will be conduct three different methods qualitative performance about OLSR and AODV to see about their implementation, and performance about those two routing method. This qualitative method that has conduct is the calculation method of the mathematical model, network simulation method, and field testing methods. The network type that have been use to this experiment is type A (using three nodes), type B (using four nodes), type C (using 5 nodes) and for testing a complex network (more than 10 nodes) will be used a network simulation QualNet. Based on the testing results, we can conclude that quantitative performance of AODV routing protocol is better than the OLSR routing protocol in a simple network (no more than 10 nodes), while the OLSR routing on complex networks (more than 10 nodes) better than AODV.

Keywords

MANET, ad-hoc network routing protocols, OLSR, AODV

1. INTRODUCTION

Information transfer is one of a major issue in information technology development. This is because one of basic purpose of development of information technologies is intended to transfer information between the parties. One of the latest developments in information transfer is the routing called MANET (Mobile Ad-hoc Network) which is used as one standard routing on the wireless world. MANET itself is divided into two methods are proactive routing method, which is represented by the Optimized Link State Routing (OLSR) and reactive routing method, which is represented by the Ad-hoc On-demand Distance Vector (AODV) [4].

In this research, will be conduct three different methods qualitative performance about OLSR and AODV to see about their implementation, and performance about those two routing method. Methods that have been used to conduct a qualitative performance are a calculation method using mathematical model, network simulation method, and field testing methods. The network type that have been use to this experiment is type A (using three nodes), type B (using four nodes), type C (using 5 nodes) and for testing a complex network (more than 10 nodes) will be used a network simulation QualNet. For testing data transfer will be done a continuously transfer data in certain numbers of packet and a certain packet size, so that later the performance between the AODV and OLSR based on this testing variables.

2. PROACTIVE AND REACTIVE ROUTING

Proactive routing (figure 1) determine the routes to some nodes in a network that has been developed so that the route will always be ready when needed. Overhead for this routing is large enough because each node must discover all existing routes in the network, thus this method will be create a relative large bandwidth consume to keep this routes keep up-to-date. But in exchange, the package transmit is become fast enough because the route is already exists. Example for this method is like Destination sequenced Distance Vector (DSDV), Optimized Link State Routing (OLSR) and GSR [3].

Meanwhile, reactive routing determines the route only if its necessary so that the overhead of Route Discovery is quite small, this method uses the mechanism of flooding (global search). But in exchange a node that will transmit a packet must wait for the discovery of a route. Examples of reactive routing instance: Dynamic Source Routing (DSR), Ad-hoc On-demand Distance Vector (AODV) and TORA [3].

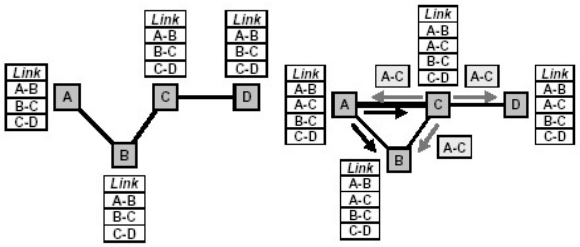


Figure 1. Example of proactive routing algorithm [5]

3. MOBILE AD-HOC NETWORK (MANET)

Mobile Ad-hoc Network (MANET) is one of ad-hoc wireless network type. MANET is a self-configuring network from a multiple mobile routers (and associated hosts) connected by wireless links [2]. Routers are free to move randomly and organize themselves dynamically so that the wireless network topology can change drastically and can not be predicted [9]

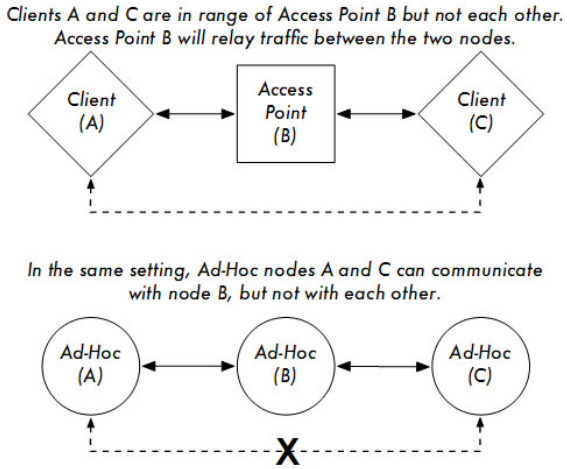


Figure 2. Ad-hoc mode and managed mode [9]

One of the lacks from an ad-hoc mode is the inability of the node to forward data packets to the third node (figure 2). If the network using an access point, even node A and node C not in each other range area, but they can still communicated each other through the access point that still within their reach area. In the ad-hoc mode, node A and C can't communicate each other because their location is out off their range area (figure 2). But with a routing protocol, the second node in the middle is inserted in the ad-hoc mode (figure 3), packet can carry data from the first node (A) to the third node (C). In this case, the second node to act as a relay to widen the reach of wireless networks (figure 4). One of the implementation of mesh routing technique is a MANET.

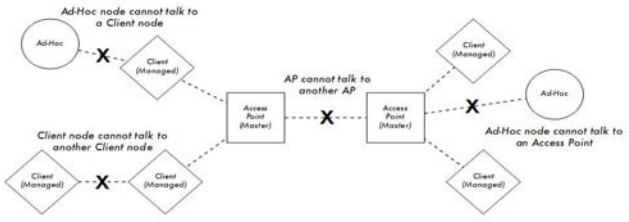


Figure 3. Master mode (access point) and client/managed mode [9]

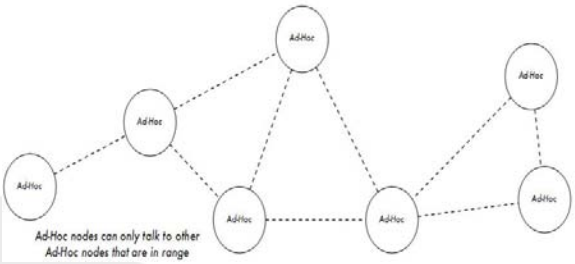


Figure 4. Ad-hoc mode [9]

4. OPTIMIZED LINK STATE ROUTING (OLSR)

Optimized Link State Routing (OLSR) is a proactive routing in mobile ad-hoc network. This protocol has the stability of link state algorithm and has the advantage with a route that's quickly available when it's needed. OLSR is an optimization of the classical link state protocol designed for wireless network usage.

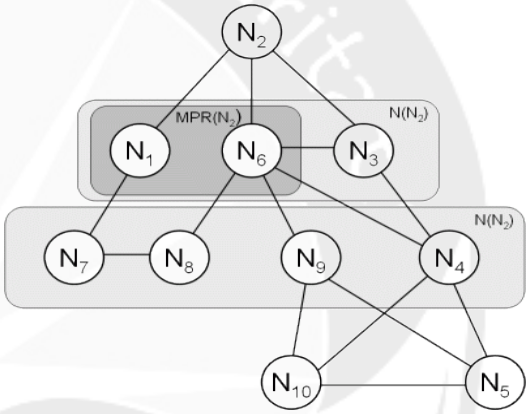


Figure 5. Examples of routing protocol OLSR [8]

Each node on the network, for example in figure 5 is node N2, will select multiple nodes in the network of his neighbors. These nodes will send packets to node N2. Neighboring nodes, namely the N1 and N6 called Multipoint relays of node N2. Node N2 chose him for the Assembly to cover all the nodes that are two hops away, for example node N7, N8, N9, and N4.

Beside that, OLSR does not require sequential message delivery. Each control message has a sequence number that otomatically increase for each message. This causes the receiver of the message can identify the latest message [6].

5. AD-HOC ON-DEMAND DISTANCE VECTOR (AODV)

On-demand Distance Vector (AODV) routing protocol is designed for ad-hoc network [1]. Which AODV can perform unicast and multicast routing. AODV is a reactive routing protocol that use on-demand-based algorithm, which means that this protocol will make the route in the network only if it's required by the source node to send a message. AODV route runs only as long as needed by the source. Additionally, AODV makes tree connecting member and the member-node multicast group.

In AODV, to find a route to destination, the source will broadcasts route request packets to the neighbor. The neighbor node will then broadcast the packet to their neighbor until it reaches the node that

has information about the node destination or until it reaches the destination node. Route request packet will be used a sequential numbers to ensure that these nodes will repply only with the latest information alone [5] [6]

When a node sends a route request to neighboring nodes, the package also store information from which the package first arrived in its routing table. This information is used to create a route back from the route request packet. AODV uses only symmetric links because the route request packets follow the route back from the route request packet. Whereby when the route reply packet transmitted back to the source (figure 2), the nodes along the route include further routes into its routing table.

The advantage from AODV is that this protocol does not create additional traffic on the communications links that already exist. This makes routing simple and does not require a lot of memory allocation for routing calculations. However, AODV needs more time to create connections, and initial communication needed to create sometimes more difficult than some other methods [7]

6. NETWORK DESIGN

To perform quantitative performance test in data transfer between MANET proactive routing protocol (OLSR) and reactive routing method (AODV), there's three types of networks (type A, B, and C) that designed for represent several type of ad-hoc wireless networks. Ranging from relatively simple to quite complicated network. Three types of this networks are as follows (figure 6, 7, and 8). Where the Laptop source will be placed on T building and laptop destination will be place on W building.

6.1 Network Structure Type A

Network Structure Type A (figure 6) builds by three wireless ad-hoc devices using three laptops. Network Structure Type A designed with the simplest structure among this three types of experimental network, so the performance of this type is expected become the best-performing network.

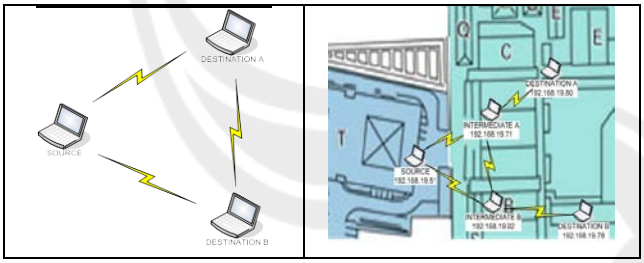


Figure 6. Network structure type A

6.2 Network Structure Type B

Network Structure Type B builds by four wireless ad-hoc devices using four laptops (Figure 8).

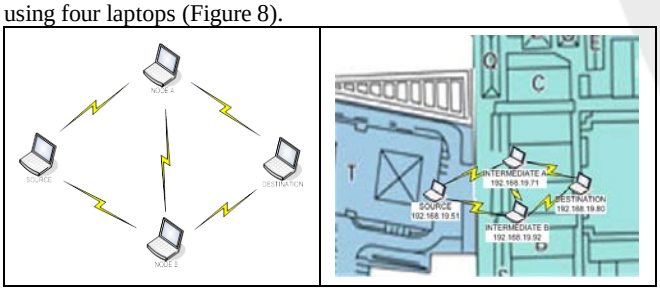


Figure 7. Network structure type B

6.3 Network Structure Type C

Network structure type C build by five wireless ad-hoc devices using five laptops. This network type is designed with the most complicated structure among another network structure types for this experiment, so this performance supposed become the worst.

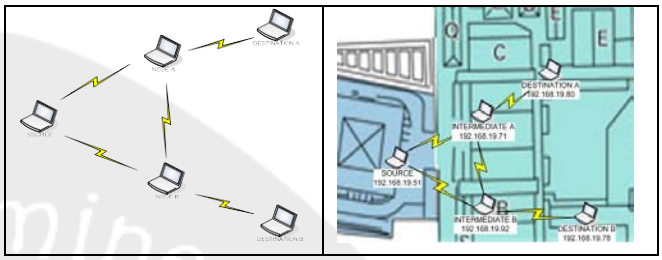


Figure 8. Network Structure Type C

7. Implementation and System Testing

Implementation and system testing for both OLSR and AODV is performed by delivery a several data packets, up to 30 packages with each package size is 512 bytes. Testing also also conducted with a large delivery of data packets from 512 bytes to 16 kilobytes. The process of comparison of results of OLSR and AODV will be based on packet delivery ratio, end to end delay, packet control ratio, path length ratio, and throughput generated by network structure design type A, B, and C.

7.1 Testing on Network Structure Type A

7.1.1 Network Test based on Amount of Packet Transmision on Network Structure Type A

On this experiment, each of network structure will be tested by a number of packets (1, 5, 10, 15, 20, 25 and 30 packets) that transmitted from source node to destination node with a packet size 512 bytes for each Packet. The simulation results from network type A using AODV routing can be seen in table 1. Meanwhile, test results against OLSR routing based can be seen in table 2

Table 1. Network structure type A testing using AODV routing based on amount of packets

Amount of Packets	Packet Delivery Ratio (%)		Delay End-to-End (milisecond)		Packet Control Ratio (O/H)	Path Length Ratio	Throughput (bit/s)
	Dest. A	Dest. B	Dest. A	Dest. B			
1	100	100	1.7	2.4	4:1	1:1	0
5	100	100	6.1	10.3	4:5	1:1	5200
10	100	100	4.7	8.6	4:10	1:1	4600
15	100	100	4.3	8.1	4:15	1:1	4400
20	100	100	4.1	7.8	4:20	1:1	4300
25	100	100	3.9	7.6	4:25	1:1	4250
30	96	96	3.8	7.5	4:30	1:1	4100

Table 2. Network structure type A testing using OLSR routing based on amount of packets

Amount of Packet	Packet Delivery Ratio (%)		Delay End-to-End (milisecond)		Packet Control Ratio (O/H)	Path Length Ratio	Throughput (bits/s)
	Dest. A	Dest. B	Dest. A	Dest. B			
1	50	50	3.3	6.6	15:1	5:3	0
5	60	60	3.4	6.7	15:5	5:3	6200
10	80	80	3.4	6.8	15:10	5:3	4700
15	86	86	3.4	6.9	15:15	5:3	4400
20	90	90	3.3	6.9	15:20	5:3	4300
25	92	92	3.3	6.9	15:25	5:3	4300
30	93	93	3.3	6.9	15:30	5:3	4200

From table 1 and 2 we can see that both the routing AODV and OLSR routing has their own superiority on different variables. However, there is a tendency that AODV routing have a better performance than OLSR routing on network structure type A.

In figure 9, shows the results of packet delivery ratio from network type A based on amount of packets, which shows that AODV routing is better than OLSR routing. In the figure 10, shows that the results delay end-to-end network type A from routing AODV is better than OLSR. Except on the first testing, it shows that the delay end-to-end on AODV is smaller than OLSR.

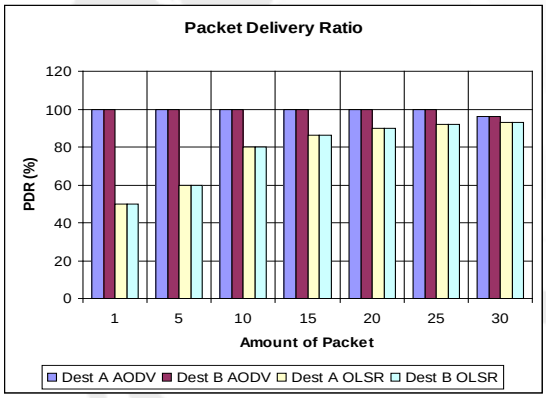


Figure 9. Packet delivery ratio testing from network type A based on amount of packets

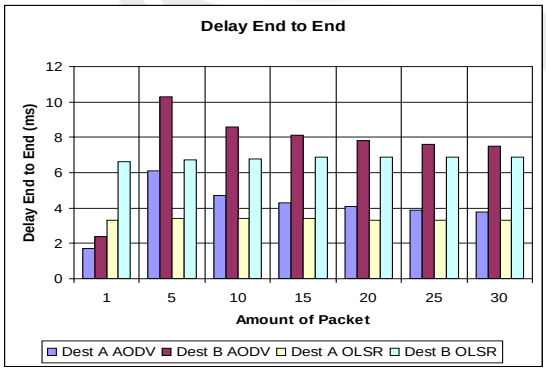


Figure 10. Delay end-to-end from network type A based on amount of packets

7.1.2 Network Test based on Size of Packet Transmision on Network Structure Type A

The parameters have been used in this testing based on the size of packets that transmitted from the source node to destination node. The size of the packets is range from 512 to 16.384 bytes, for each test the source will send five packets to destination. The testing using AODV routing on Network Structure type A can be seen in Table 3, while for OLSR in Table 4.

From table 3 and 4 we can see that both the routing AODV and OLSR routing has their own superiority on different variables. However, there is a tendency that AODV routing have a better performance than OLSR routing on network structure type A.

Table 3. Network structure type A testing using AODV routing based on the size of packets

Packet Size (bytes)	Packet Delivery Ratio (%)		Delay End-to-End (milisecond)		Packet Control Ratio (O/H)	Path Length Ratio	Throughput (bits/s)
	Dest. A	Dest. B	Dest. A	Dest. B			
512	100	100	6.1	10.3	4:5	1:1	5200
1024	100	100	8.1	14.1	4:5	1:1	10300
1536	100	100	10.2	18.2	4:5	1:1	15500
2048	100	100	14.2	25.2	4:5	1:1	20600
3072	100	100	18.2	33.8	4:5	1:1	31000
4096	100	100	24	44.8	4:5	1:1	41000
6144	100	100	34	64	4:5	1:1	62000
8192	100	100	43	84	4:5	1:1	82500
12288	100	100	64	124	4:5	1:1	124000
16384	100	100	82	162	4:5	1:1	165000

Table 4. Network structure type A testing using OLSR routing based on the size of packets

Packet Size (bytes)	Packet Delivery Ratio (%)		Delay End-to-End (milisecond)		Packet Control Ratio (O/H)	Path Length Ratio	Throughput (bits/s)
	Dest. A	Dest. B	Dest. A	Dest. B			
512	60	60	3.4	6.7	15:5	5:3	6200
1024	60	60	7	10.9	15:5	5:3	12200
1536	60	60	7.5	14.9	15:5	5:3	18500
2048	60	60	11	21.9	15:5	5:3	24600
3072	60	60	15	30.1	15:5	5:3	37000
4096	60	60	20.2	42	15:5	5:3	49000
6144	60	60	30.2	60.4	15:5	5:3	74000
8192	60	60	39	80	15:5	5:3	98000
12288	60	60	60	120	15:5	5:3	148000
16384	60	60	78	158	15:5	5:3	198000

On packet delivery ratio and delay end to end testing AODV has a greater result than OLSR. But, on control packet and path length ratio AODV has a smaller result than OLSR, the same result is also happened on testing using amount of packets. On throughput both of AODV and OLSR have a quite same result.

7.2 Testing on Network Structure Type B

7.2.1 Network Test based on Amount of Packet Transmision on Network Structure Type B

Testing network structure type B based on the amount of packets is shown in table 5 (for AODV) and table 6 (for OLSR). Where is seen that the packet delivery ratio in AODV is greater than in OLSR, as well as the delay of end-to-end and control packet. While the ratio for the path length was found that AODV is smaller than OLSR. Meanwhile, the network throughput for type B shows that in almost all the testing, both AODV and OLSR have a similar result except in fifth test shows that the throughput of OLSR which is greater than AODV.

Table 5. Network structure type B testing using AODV routing based on amount of packet

Amount of Packet	Packet Delivery Ratio (%)	Delay End-to-End (milisecond)	Packet Control Ratio (O/H)	Path Length Ratio	Throughput (bits/s)
1	100	13.0	4:1	1:1	0
5	100	31.8	4:5	1:1	5300
10	100	19.5	4:10	1:1	4600
15	100	15.4	4:15	1:1	4400
20	100	13.4	4:20	1:1	4300
25	100	12.2	4:25	1:1	4250
30	96	11.5	4:30	1:1	4200

Table 6. Network structure type B testing using OLSR routing based on amount of packet

Amount of Packet	Packet Delivery Ratio (%)	Delay End-to-End (milisecond)	Packet Control Ratio (O/H)	Path Length Ratio	Throughput (bits/s)
1	50	7.1	29:1	14:4	0
5	60	7.1	29:5	14:4	6200
10	80	7.1	29:10	14:4	4700
15	86	7.1	29:15	14:4	4400
20	90	7.1	29:20	14:4	4300
25	92	7.1	29:25	14:4	4300
30	93	7.1	29:30	14:4	4300

7.2.2 Network Test based on Size of Packet Transmision on Network Structure Type B

Testing network structure type B based on the size of packets is shown in table 7 (for AODV) and table 8 (for OLSR). Which, the result on packet delivery ratio, delay end-to-end, and control packet ratio showed that AODV packet has a greater result than OLSR. While for path length ratio and throughput AODV result is smaller than OLSR.

Table 7. Network structure type B testing using AODV routing based on size of packets

Packet Size (bytes)	Packet Delivery Ratio (%)	Delay End-to-End (milisecond)	Packet Control Ratio (O/H)	Path Length Ratio	Throughput (bits/s)
512	100	31.8	2:5	1:1	5300
1024	100	35.8	2:5	1:1	10600
1536	100	40	2:5	1:1	15800
2048	100	47	2:5	1:1	21200
3072	100	55	2:5	1:1	31800
4096	100	66.5	2:5	1:1	42000
6144	100	86	2:5	1:1	63000
8192	100	105	2:5	1:1	84000
12288	100	145	2:5	1:1	127000
16384	100	184	2:5	1:1	169000

Table 8. Network structure type B testing using OLSR routing based on size of packets

Packet Size (bytes)	Packet Delivery Ratio (%)	Delay End-to-End (milisecond)	Packet Control Ratio (O/H)	Path Length Ratio	Throughput (bits/s)
512	60	7.1	29:5	14:4	6200
1024	60	11.2	29:5	14:4	12300
1536	60	15.2	29:5	14:4	18500
2048	60	22.8	29:5	14:4	24600
3072	60	31	29:5	14:4	37000
4096	60	42	29:5	14:4	49000
6144	60	62	29:5	14:4	74000
8192	60	81	29:5	14:4	98000
12288	60	122	29:5	14:4	148000
16384	60	159	29:5	14:4	198000

7.3 Testing on Network Structure Type C

7.3.1 Network Test based on Amount of Packet Transmision on Network Structure Type C

Testing network structure type C based on the amount of packets is shown in table 9 (for AODV) and table 10 (for OLSR). Where is seen that the packet delivery ratio in AODV is greater than in OLSR. While for delay end-to-end testing almost all the testing packages (except for 25 and 30 packets testing), delay end-to-end on AODV is greater than OLSR. Meanwhile, on the packet control ratio and path length ratio, AODV test result is less than OLSR. Meanwhile, for throughput testing showed that the throughput at AODV and OLSR relatively the same result, except for the fifth test, which the throughput of OLSR packet is smaller than the AODV.

Table 9. The simulation results of AODV Routing Type C network based amount of packet

Amount of Packet	Packet Delivery Ratio (%)		Delay End-to-End (milisecond)		Packet Control Ratio (O/H)	Path Length Ratio	Throughput (bits/s)
	Dest. A	Dest. B	Dest. A	Dest. B			
1	100	100	114	124	4:1	1:1	0
5	100	100	34	31	4:5	1:1	5250
10	100	100	23.8	20.8	4:10	1:1	4600
15	100	100	20.8	17	4:15	1:1	4400
20	100	100	19	15	4:20	1:1	4300
25	100	100	18.1	13.6	4:25	1:1	4300
30	96	96	17.6	12.9	4:30	1:1	4200

Table 10. The simulation results of OLSR Routing Type C network based amount of packet

Amount of Packet	Packet Delivery Ratio (%)		Delay End-to-End (milisecond)		Packet Control Ratio (O/H)	Path Length Ratio	Throughput (bits/s)
	Dest. A	Dest. B	Dest. A	Dest. B			
1	50	0	7.4	0	34:1	19:5	0
5	60	20	7.2	14.5	34:5	19:5	3100
10	80	60	10.2	12.5	34:10	19:5	4600
15	86	73	10.6	12.6	34:15	19:5	4450
20	90	80	10.6	12.8	34:20	19:5	4300
25	92	84	10.4	13.1	34:25	19:5	4250
30	93	86	10.3	13.2	34:30	19:5	4200

7.3.2 Network Test Based on Size of Packet Transmision on Network Structure Type C

Testing network structure type C based on size of packets is shown in table 11 (for AODV) and table 12 (for OLSR). Where is seen that the packet delivery ratio, delay end-to-end, and throughput in AODV is greater than in OLSR. While for packet control ratio dan path length ratio AODV is smaller than OLSR.

Table 11. The simulation results of AODV routing type C network based on size of packet

Packet Size (bytes)	Packet Delivery Ratio (%)		Delay End-to-End (milisecond)		Packet Control Ratio (O/H)	Path Length Ratio	Throughput (bits/s)
	Dest. A	Dest. B	Dest. A	Dest. B			
512	100	100	34	31	4:5	1:1	5250
768	100	100	39	33.6	4:5	1:1	7900
1024	100	100	42	36	4:5	1:1	10500
1280	100	100	45	38	4:5	1:1	13100
1536	100	100	48	42	4:5	1:1	15800
1792	100	100	52.5	44	4:5	1:1	18500
2048	100	100	62	57	4:5	1:1	21000
2304	100	100	65	59	4:5	1:1	23700
2560	100	100	68	62	4:5	1:1	26200
2816	100	100	72.6	67	4:5	1:1	29000

Table 12. The simulation results of OLSR routing type C network based on size of packet

Packet Size (bytes)	Packet Delivery Ratio (%)		Delay End-to-End (milisecond)		Packet Control Ratio (O/H)	Path Length Ratio	Throughput (bits/s)
	Dest. A	Dest. B	Dest. A	Dest. B			
512	60	20	7.2	14.5	34:5	19:5	3100
768	60	20	9.3	18.6	34:5	19:5	4625
1024	60	20	11.2	22.8	34:5	19:5	6150
1280	60	20	13.4	26.8	34:5	19:5	7700
1536	60	20	15.5	31	34:5	19:5	9250
1792	60	20	17.7	35	34:5	19:5	10800
2048	60	20	25.5	45.1	34:5	19:5	12200
2304	60	20	27.5	49	34:5	19:5	13900
2560	60	20	29.5	53.5	34:5	19:5	15300
2816	60	20	32	57.5	34:5	19:5	16900

7.4 Testing Results on Complex Networks

To test a complex network that consists of more than 10 nodes will be tested using simulation software Qualnet, as been seen on figure 11. The result on testing packet delivery ratio for a complex network (more than 10 nodes) in the static condition can be seen on table 13 and 14. The result shows that OLSR routing has better quantitative performance compared with AODV.

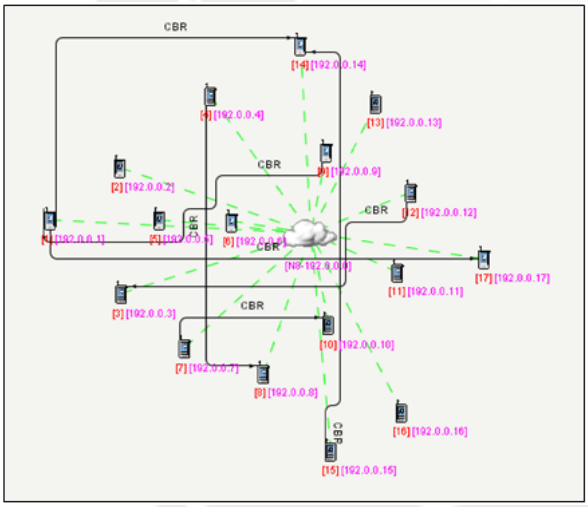


Figure 11. Complex network design using qualnet simulation software

Table 13. AODV Packet Delivery Ratio Testing on Complex Network

Packet Size (bytes)	Packet Delivery Ratio (%)						
	node 1-17	node 1-14	node 15-14	node 7-10	node 4-8	node 12-3	node 9-1
512	40	40	60	60	40	40	20
768	40	40	40	60	40	40	20
1024	40	40	40	60	40	40	20
1280	40	40	40	60	40	40	20
1536	40	40	40	60	40	40	20

Table 14. OLSR packet delivery ratio testing on complex network

Packet Size (bytes)	Packet Delivery Ratio (%)						
	node 1-17	node 1-14	node 15-14	node 7-10	node 4-8	node 12-3	node 9-1
512	100	80	100	100	80	100	100
768	100	80	80	100	100	80	100
1024	80	60	100	100	100	100	100
1280	100	60	100	100	100	100	100
1536	60	100	60	100	100	80	80

8. CONCLUSION

The conclusion that can be obtained based on the design and testing is that there's a tendency that the quantitative performance from AODV routing protocol is better than OLSR in a network that less complex (less than 10 nodes) either on the network type A, B and C. But for a complex network (more than 10 nodes) there's a tendency that quantitative performance from OLSR routing protocol is better than AODV.

9. REFERENCES

[1] AODV description. (n.d.).http://moment.cs.ucsb.edu/AODV/aodv.html

[2] Forouzan, Behrouz, A. (2004). Data Communications and Networking 3rd Edition. New York: McGraw-Hill.

[3] Khetrapal, Ankur. (2003). Routing Techniques for Mobile Ad-Hoc Networks Classification and Qualitative/Quantitative Analysis. Computer Engineering Department Delhi University.

[4] Misra, Padmini. (1999). Routing Protocols for Ad Hoc Mobile Wireless Networks.

[5] Raghavan, Sudarshan Narasimha. (2003). The Terminal Node Controlled Routing Protocol for Mobile Ad Hoc Networks.

[6] University of Luxembourg, SECAN-Lab (2004). Optimized Link State Routing Protocol.

[7] Wikipedia. (2010). Ad-hoc On-Demand Distance Vector Routing. http://en.wikipedia.org/wiki/AODV

[8] Wikipedia. (2010). Optimized Link State Routing Protocol. http://en.wikipedia.org/wiki/OLSR

[9] Wikipedia. (2010). Mobile Ad-hoc Network. http://en.wikipedia.org/wiki/MANET

Spatial Rain Rate Measurement to Simulation Colour Noise Communication Channel Modeling for Millimeter Wave In Mataram

Made Sutha Yadnya
Universitas Mataram, Mataram
msyadnya@unram.ac.id

Gamantyo Hendrantoro
Institut Teknologi Sepuluh November, Surabaya
gamantyo@ee.its.ac.id

ABSTRACT
This paper presents an initial result of research from rain rate measurements in Mataram Lombok Indonesia. Rain rate is intruder for channel communication espesially in Local Multipoint Distribution Service transmission at 30 GHz. Channel charactization in order to understand and mitigate the problem. Parameter statistic of rain rate to proceces prediction used design and evaluation LMDS system. Data rain rate from urban mensureament made channel noise used union to generate signal rain rate. Mean and variance for the check normal data used large-scale fading. Large-scale fading is lognormal distribute. Noise in channel from rain rate made classification in colour noise.

Keyword
Rain rate, Channel communication, Noise

1. INTRODUCTION
This paper deals with the estimation of noise correla- tions along an array of sensors. The aim of this paper is not to add a supplementary method for spectral (spatial) analysis; it is to present feasible methods for estimation of noise correlations in the presence of point sources (ships, etc.). Direct ARMA modeling of sensor outputs (sources plus noise) is a way to consider noise with arbitrary correlations, but it is not well suited to array processing, the main difficulty being due to high-order modeling (which is necessary for a great number of sources). The separation of the space of observations (sensor outputs) and sources subspaces) is a better way for noise correlation estimation. The main advantage of these approaches relies upon the low-order model of noise. The adequacy of an AR(MA) modeling of noise is, obviously, a crucial point and will be considered later. It is also important to consider the improvements obtained by means of the noise correlation estimated [1].

We present principally two types of methods for the estimation of noise parameters (estimation of the ARMA coefficients). The first is related to the calculation of the likelihood of whitened observations (by means of ARMA noise modeling)

Indonesia needs broadband communication next generation, because newera multimedia telecommunication in the other place has used (LDMS). Wireless communication used 30 GHz has installed for high-speed communication, the otherwise rain rate has intruder to this communication. This condition challenging to design and evaluation system for simulation channel wireless impact rain rate [2].

2. RAIN RATE MODELING
Rain rate is random condition. Random have inferential statistics, than it is interested in making an inference about the characteristics of a population through information obtained in a subset called sample. To make an inference about a population parameter (characteristic), population draws a random sample from the population. Random sampling procedure is one in which every possible sample of *n* observations from the population is equally likely to occur, suppose can select a sample of size *n* from a population of size N, and attempt to make an inference about the population mean by drawing a sample from the population and calculating the sample mean. Exsample for sample is $x_1, x_2, ..., x_n$, are n independent random variables from a population with mean μ and variance σ[4]. Then the sum or average of those variables will be approximately normal with mean μ and variance σ2/n as the sample size becomes large [3].

The lognormal distribution is an asymmetric distribution with interesting applications for modeling the probability distributions of stock and other asset prices. A continuous random variable x follows a lognormal distribution if its natural logarithm, ln(x), follows a normal distribution. Distribution can also say that if the natural log of a random variable, ln(x), follows a normal distribution, the random variable, x, follows a lognormal distribution. Interesting observations about the lognormal distribution is the lognormal distribution is asymmetric (skewed to the right). Recall that the exponential and logarithmic functions mirror each other.

Model rain rate at line millimeter wave is assumes distribution lognormal for rain rate with statistical parameters and function of auto covariance is derivable from mean, standard deviation, and function of autocovariance from value logarithmic rain rate[2].

Normal N(μ, σ) (1a)

Generate Lognormal:
LogNormal LN*(m,v)* (1b)

Mean generate signal :

$$m = e^{\mu + \frac{\sigma^2}{2}} = e^{2\mu + \sigma^2} \quad (2)$$

Variance generate:

$$v = e^{2\mu + \sigma^2}(e^{\sigma^2} - 1) \quad (3)$$

Variance generate :

$$v = m^2(e^{\sigma^2} - 1) \quad (3a)$$

Exponential generate:

$$e^{\sigma^2} = \frac{v}{m} + 1 \quad (4)$$

Variance generate:

$$\sigma^2 = \ln(\frac{v}{m} + 1) \quad (5)$$

Mean generate :

$$m = e^{\mu+\frac{\sigma^2}{2}} \quad (6)$$

Mean generate :

$$\mu + \frac{\sigma^2}{2} = \ln(m) \quad (7)$$

Mean final generate :

$$\mu = \ln(m) - \frac{\sigma^2}{2} \quad (7a)$$

Calculate for expentation y from exponential condition lognormal distribution :

$$y = e^{\left(\mu+0.50\sigma^2\right)} \quad (8)$$

From regression linier found formula :

$$\ln(x) = y \Leftrightarrow e^y = x \quad (9)$$

Probability lognormal dirtibution formula is :

$$p_{IN}(x) = \frac{1}{\sigma\sqrt{2\pi}}\frac{1}{x}\exp\left[-\frac{1}{2}\left(\frac{\ln(x)-\mu}{\sigma}\right)^2\right] \quad (10)$$

2.1. Auto Regressive (AR) Modeling

Stochastic modeling is assumed rain rate r mm/hr be wide sense stationary and distribution lognormal, hence this also express that specific damping of rain γ (dB/km) along the length of radio line (link) also distribution lognormal and stationery[2]. So η = in α logarithm natural from rain rate normal distribution will with parameter taken away from by measurement of field. Parameter is median from rain damping α_m (equivalent with μ_η from η) and deviation σ_η from η [2].

Then Assumption is function of autocorrelation $\phi_R(\tau)$ from rain damping known or measurement directly from data yielded. For function of autocorrelation normalization from r rainfall having distribution lognormal, where τ it is time delay, hence function of obtainable autocovarian. Evocation procedure of rainfall looks like evocation of Raleigh fading. A normal distribution series with mean zero and η_o (k) = η_o (kτ) where k is integer and τ It is sampling time can be awakene in recursive[3]:

$$\eta_0(k) = -\sum_{n=1}^{M} a(n)\eta_0(k-n) + c\,g(k) \quad (11)$$

where *a*(n) be coefficient AR, n = 1 ,..., M, M is number of orders from process depended from maximum delay, *g*(*k*) be random series number Gaussian mean 0 and variant 1 awakened with computer[4], c is factor is donation of deviation standard from series noise *cg*(*k*)[5]. With getting of series η_o (*k*) hence series r(*k*) obtained with equation:

$$\mathrm{r}(k) = \exp\left(\eta_0(k) + \mu_\eta\right) \quad (12)$$

2.2. Moving Average (MA) Modeling

Coefficient MA is got with step of first of all estimate input sample (in this case rainfall result of measurement). After estimating parameter AR high order by the way of looking for inverse from a filter AR like the one has done with method circular lattice, which written down with equation:

$$h_i(n) = \delta(n) + \sum_{k=1}^{p} a_k \delta(n-k) \quad (13)$$

Where a_k is coefficient AR, p is order from series AR, this is filter inverse AR. Filter AR in transfer function channel assume IIR filter. After estimating then awakens estimation of input sample from data with rainfall data convolute with inverse from filter AR. The estimation can be awakene with equation:

$$e(n) = y(n) * h_i(n) \quad (14)$$

2.3. Auto Regressive Moving Average (ARMA) Modeling

Auto Regressive Moving Average is fusion from AR an MA, condition to evocation (generate) of rain rate with this ARMA process made equation shown in:

$$y(n) = -\sum_{i=1}^{p} a(i)y(n-i) + \sum_{j=0}^{q} b(j)v(n-j) \quad (15)$$

3. SIMULATION

Data measurement processed by Matlab. Focus data at November 2009 than use to simulation for precdiction and estimation. Rain rate simulation in convective condition, this condition fade slope very deep and attenuation from rain rate should be stop. Figure 1 probabiliti distribution rain rate condition. Figure 2 is spatial measurement from 3 rain gauge compire 2 rain gauge BMKG (Badan Metreologi Klimatologi dan Geofisika Mataram. Figure 3 normalitation from 5 rain gauge use to model variation moving rain rate. Figure 4 Rain gauge measureament condition rain rate in spatial. Figure 5 Resault rain gauge measureament made correlation used analized moving rain rate in Mataram.

4. DISCUSSSION

Rain is intruder for signal transmission and become big problem in telecommunications area of technology wireless. For frequency 30 GHz, millimeter having weakness, because the wave is very short is order to achieve good performance, which is good needs design anti fading. Variation time and spatial in Mataram have fade slope very high to mitigate.

5. CONCLUSION

Rain rate from data measurement assume about rain rate is lognormal distribution used to estimation spectral of rain rate considers temporal characteristics (time domain). The results suggest more accurate prediction channel provides an evaluation for a worse scenario than the ITU-R method, and hence is more recommend for use in the design millimeter wave communication system for tropical maritime region. Data rain rate from urban mensureament made channel noise used union to generate signal rain rate. Mean and variance for the check normal data used large-scale fading. Large-scale fading is lognormal distribute. Noise in channel from rain rate made classification in colour noise. Expectation from colour noise to mitigate rainrate condition is solution to design LMDS.

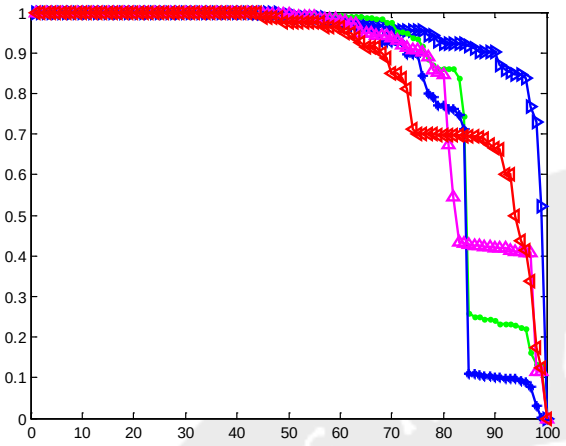


Fig.1 Simulation Distribution Rain Rate in Mataram.

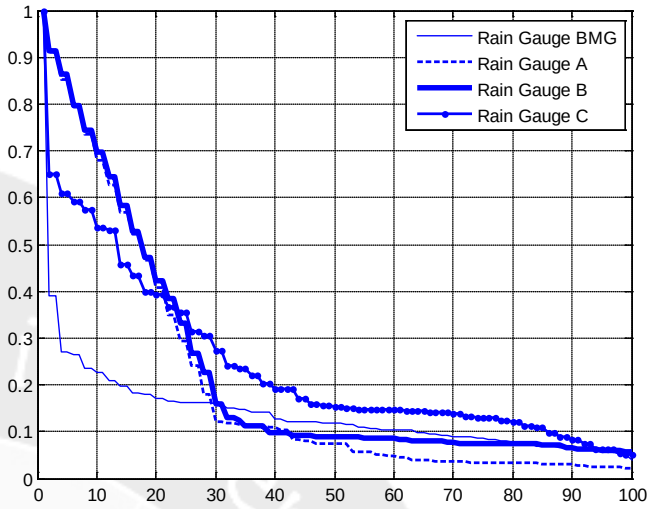


Fig.3 Simulation data rain rate variation spatial and time.

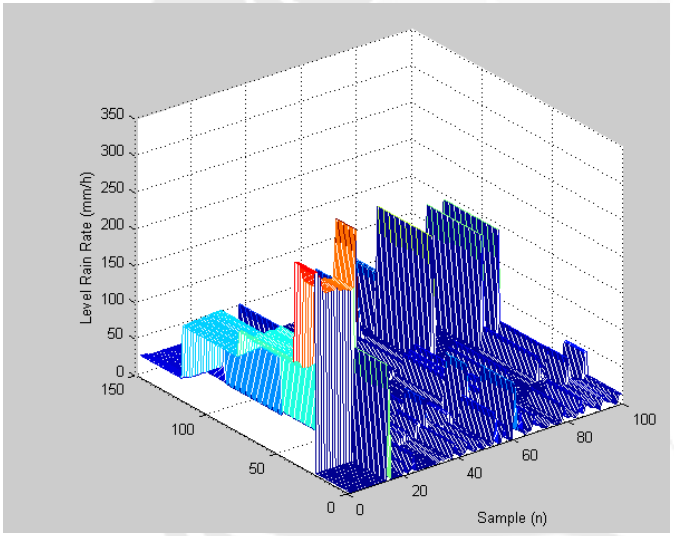


Fig.2 Simulation 3D Rain rate variation spatial and time

Fig.4 Rain gauge measureament

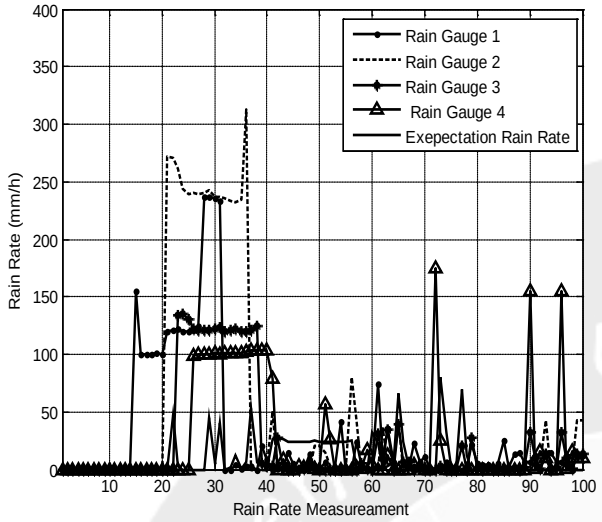


Fig.5 Correlation Spatial Time Rain Rate from Rain gauge measureament

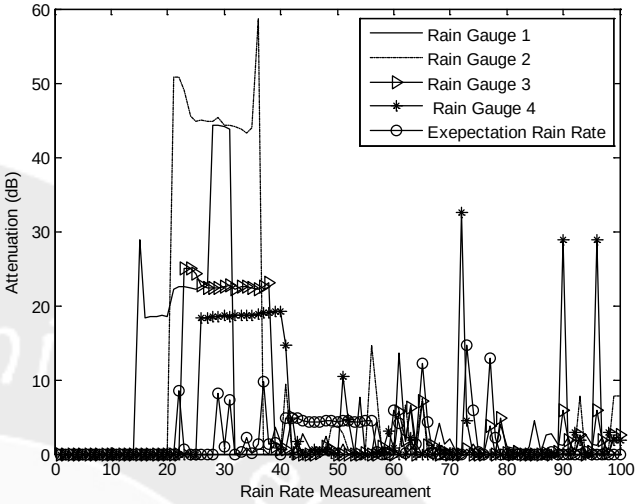


Fig.6 Correlation Spatial Time in attenuation Large-Scale

6. ACKNOWLEDGMENT

The reported research has been supported financially by JICA with project of PREDICT-ITS and by DP2M Dikti of Minister Education National award for Kompetensi Research 2009 batch I and Kompetensi Research 2010 batch II .

7. REFRENCES

[1] Cadre P. L 1989, " Parametric Methods for Spatial Signal Processing in the Presence of Unknown Colored Noise Fields". IEEE 1989 TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING. VOL. 37. NO. 7. JULY 1989.

[2] G. Hendrantoro, 2004, "*An Autoregressive Model for Simulation of Time-Varying Rain Rate*", ANTEM 2004

[3] Yadnya M.S., A. Mauludiyanto, A. Wijayanti, Mahmudah, Muriani, G.Hendrantoro,'*Simulation of Rain Rate and Attenuation in Indonesia for Evaluation of Millimeter-wave Wireless System Transmission*", ICSIIT 2007

[4] Lemorton.J.O, 2005, " Channel Models for Simulation" Cost 280 third International Workshop.

[5] Yadnya M.S., A. Mauludiyanto, G.Hendrantoro,"*Simulation of Rain Rate for Channel Communication in Surabaya*", ICAST 13-14 March 2008 Japan.

[6] Yadnya M.S, A. Mauludiyanto, G.Hendrantoro, "*Statistical of Rain Rate for Wireless Channel Communication in Surabaya*",WOCN 5-7 May 2008

[7] Yadnya M.S, A. Mauludiyanto, G.Hendrantoro, "Akaike Information Criteria Application to Stationary and Nonstationary Rainfalls for Wireless Communication Channel in Surabaya", ICTS 5 August 2008.

[8] Yadnya M.S, A. Mauludiyanto, G.Hendrantoro, "*Statistical of Rain Rate for Wireless Channel Communication in Surabaya*", WOCN 5-7 May 2008 Surabaya-Indonesia, IEEE, ISSN 978-1-4244-1980-7-08. chap III. pp 1-5.

[9] Yadnya M.S, A. Mauludiyanto, G.Hendrantoro, "Akaike Information Criteria Application to Stationary and Nonstationary Rainfalls for Wireless Communication Channel in Surabaya", ICTS 5 August 2008. ISSN 1858-1633 , pp 292-299.

[10] Yadnya M.S, A. Mauludiyanto, G.Hendrantoro, "*Lognormal Distribution from Rain Rate Measurement to Simulation Communication Channel Modeling for Millimeter Wave in Surabaya"*. Kumamoto Forum 2008, sie 03. pp 12-13.

[11] K.Morita & Higuti, 1976, "Prediction Method of Rain Attenuation Distribution of micro-millimeter waves ", Rev Electr.communication Lab vol 24, no 7-8, pp 651-688.

The Effect of Maximum Allocation Model in Differentiated Service-Aware MPLS-TE

Bayu Erfianto
Faculty of Informatics
TELKOM Institute of Technology
Jln. Telekomunikasi Dayeuhkolot 1,
Bandung 40257 Indonesia
erf@ittelkom.ac.id

ABSTRACT
Bandwidth Constraint Model is a key component in Differentiated Service Aware MPLS traffic engineering. DSTE supports classes and allows constraint-based routing. DSTE enhance the ability of MPLS TE bandwidth reservation on a link based on per class definition. Maximum Allocation Model is one of the Bandwidth Constraint Model, which is defined as one-to-one relationship with the defined Class Type. In this paper, we evaluate the performance of DSTE where Maximum Allocation Model (DSTE MAM) is used as bandwidth constraint. The performance is also compared to original MPLS TE. As the result, MAM can improve network performance significantly in term of throughput, end-to-end delay, packet loss and link utilization.

Keywords
DiffServ, Bandwidth Constraint Model, MPLS-TE, DS-TE

1. INTRODUCTION
DiffServ, or Differentiated Service, allocates bandwidth and network resource based-on classes of traffic. The concept of DiffServ itself is to distinguish IP packet based-on Differentiated Service Code Point[11]. Hence, the focus if DiffServ is focused on forwarding plane instead of end-to-end path to guarantee IP packet delivery [12].

The limitation of conventional DiffServ is about when the fish problem exist the network [5][6]. Fish problem in DiffServ causes unbalanced traffic distribution in defined network path. If a path is overloaded, thus the traffic drop rate will become more higher. The fish problem in DiffServ network is illustrated in Figure 1. It is known that SPF (Shortest Path First) is formed following LER-LSR_1-LSR_2-LER, since that path contains the lowest network costs (such as RTT).

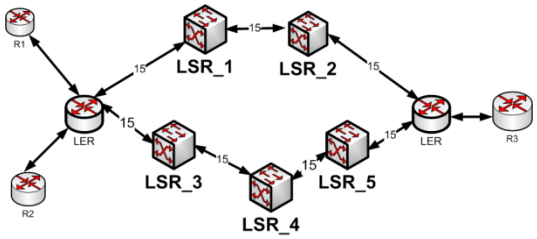


Figure 1. Fish Problem in DiffServ Network

Bandwidth Constraint Model propose by [3] is used to avoid the fish problem if DiffServ network. Actually, Bandwidth Constraint Model is a policy that defines a mechanism to allocate bandwidth regarding the different classes of traffic. This model can also be extended to DiffServ-Aware MPLS-TE (DS-TE), such as proposed in [4].

In this paper, we observe the effect of Maximum Allocation Model to the multimedia application in simulated DiffServ-Aware MPLS TE (DS-TE). We observe that DS-TE can handle the bottleneck at routers. The theoretical framework of [4] is also proved in this paper, and it is shown that no matter how congested the router, the traffic that carries multimedia data with Bandwidth Constraint Model suffers with the constant throughput, low delay, low packet drop, and optimum usage of the existing bandwidth.

- Throughput and End to End delay are observed to see the capability od DS-TE in defining a path from source to destination node
- Paket Loss is observed to see the DS-TE to handle congested in the network
- Bandwidth usage is observed to see the capability of Bandwidth Constraint Model to allocate the required bandwidth for every class of traffic with the optimum usage

This paper is the extended work of [16]. In their work, Deni et.all developed a simulation framework to set up the DS-TE simulation based on NS-2 network simulator, while the MAM as the bandwidth constraint is not taken into account.

2. DIFFERENTIATED SERVICE-AWARE MPLS-TRAFFIC ENGINEERING
MPLS Traffic Engineering (aka MPLS-TE) is defined to minimize congestion and to increase network performance [1][7][11][12]. In MPLS-TE, the routing mechanism is redefined and to allow traffic resources flow efficiently [12]. This can avoid bottleneck and network congestion, which will decrease the performance in term of jitter, delay and bandwidth utilization. MPLS-TE provides Explicit Routing through an LSP. To build an LSP, MPLS-TE involves RSVP mechanism, where ingress Label Switch Router (LER) explicitly defines Label Switch Path (LSP) to the egress LER through the intermediate core routers in between [13].

MPLS TE also extends the ordinary of MPLS routing to support Constraint Based Routing [3][6]. Instead of using IGP with single metric for routing definition, Constraint Based Routing has the

ability to calculate more details about the routing information based on the bandwidth constraint and traffic resource flowing into the network [3][6].

Nowadays, MPLS TE is combined with DiffServ TE to provide better performance by utilizing available bandwidth optimally. Such combined technique is known as Differentiated Service-Aware MPLS TE or DS-TE [4]. DS-TE provides more details control mechanism to minimize congestion in the network. To achieve this goal, DS-TE has a bit modification to support class of traffic definition to allow constraint based routing carried out in the path definition. The addition of this routing control mechanism helps DS-TE to properly control the portion of different classes of traffic flowing in the network. Basically DS-TE is carried out as control plane; however DS-TE is still use ordinary DiffServ mechanism as its forwarding plane. By means of DS-TE, it is possible to define more classes of traffic with different bandwidth allocation for each of class [4][6].

DS-TE uses Class-Type (CT) concept, which is to allocate bandwidth on each traffic class, constraint based routing, and admission control. A DS-TE network may use up to 8 CT (CT0 to CT7), which still supports priority on LSP. An LSP in DS-TE may have different priority regardless CT definition. CT in DS-TE is similar to the concept of Per Hop Behavior (PHB) and Per Hop Scheduling Class (PSC) in ordinary DiffServ. Hence, the flexible mapping between CT and PSC is possible in DS-TE [1][6].

DS-TE includes a new object definition CLASSTYPE RSVP. This object specifies CT definition related to LSP, which is defined with value 1-7. However, regarding this new object definition, a DS-TE node has to support such new object, which is inserted in the path message.

The set of Bandwidth Constraint (BC) defines policy used by a node to allocate bandwidth to CTs. For each of set, it may contain up to 9 BCs. Hence, when a DS-TE node identifies new LSP on the link, such node will use BC policy to update the unreserved bandwidth for every DS-TE class[1]. In this paper, we use Maximum Allocation Model (MAM)[4] as BC for DS-TE network, since MAM is easy to implement in the system compared to BC with Russian Doll Model (RDM). MAM is defined as follows:

- Max BC = Max Class-Type = 8
- For every c, $0 \leq c \leq$ (MaxCT-1) :
- Reserved(CTc) $\leq$ BCc $\leq$ Max-Reservable-Bandwidth
- SUM (Reserved(CTc)) $\leq$ Max-Reservable-Bandwidth
- where SUM for all c is $0 \leq c \leq$ (MaxCT-1)

Figure 2 shows a set of BC using MAM in DS-TE. In that figure, BC0 limits the bandwidth CT0 up to 15 % from maximum reserveable bandwidth; BC1 limits CT1 up to 50%, and BC2 limits CT2 up to 10%. The feature of BC in DS-TE allow each CT to receive the portion of its bandwidth definition without preemption mechanism[1].

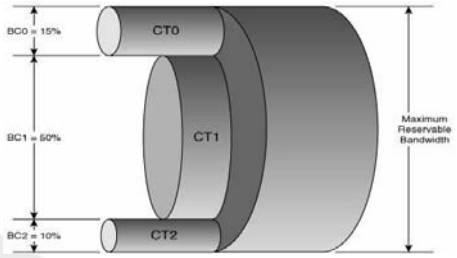


Figure 2. MAM Bandwidth Constraint Model

Basically, DS-TE uses bandwidth constraint model as a control plane to allow DS-TE to choose the LSP efficiently regarding the availability of the bandwidth. However, there exist another feature implemented in MPLS-TE which is relatively more efficient than DSTE. This feature if known as Explicit Routing. By means of this explicit routing, it is possible to allocate one flow on one LSP, because explicit routing can be defined manually by network administrator on each of MPLS TE network.

In DS-TE, one LSP is still possible to carry one or two traffic flows. However, the advantage of defining LSP in DS-TE is no network administrator intervention involved. Hence compared to MPLS-TE, the path of LSP in DS-TE can be determined automatically.

3. SIMULATION DESIGN

In this paper, a simulation model is developed based on scenario in [10]. However, in this paper, a MAM model and performance analysis is explained. The performance measure of interest in this paper covers throughput, end-to-end delay, packet loss, and bandwidth utilization.

3.1 Simulation Model

The simulation of Bandwidth Constraint Model in DS-TE Network can be performed in Network Simulator 2 (NS-2). The topology used for simulation purpose in this paper is depicted in Figure 3. Such simulated topology consists of link, node, and linkstate protocol. To meet the requirement of DS-TE model, we modify the network link hence it can provide bandwidth allocation for each of traffic flowing into the network.

Another parameter defined in DS-TE link is Admission Control to handle bandwidth request from each class of traffic. The nodes used in this simulation are defined as standard router and MPLS capable router. Linkstate protocol is further used to exchange the information of link status.

From Figure 3, it is assumed that all nodes are connected by means of 10 mbps link. From this, the Bandwidth Constraint is further defined as follows.

50% is allocated for BC1. We define that BC1 corresponds to Class Type 1 (CT1), which is allocated to carry VoIP traffic. Hence, based-on this BC1 the routing path or LSP of VoIP traffic flow is governed.

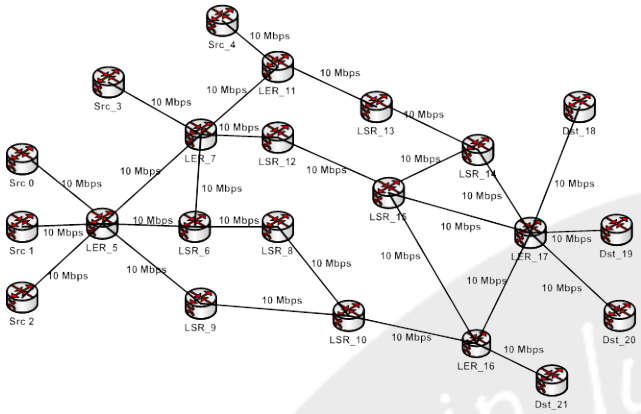


Figure 3. DS-TE Topology used in Simulation

30% is allocated for BC2. We define that BC2 corresponds to Class Type 2 (CT2), which is allocated to carry data traffic. Hence, based-on this BC1 the routing path or LSP of data traffic flow is governed.

20% is allocated for BC3. We define that BC3 corresponds to Class Type 3 (CT3), which is allocated to carry Best Effort (BE) traffic. However, BE traffic is not to be included in QoS definition, and the 20% of allocated bandwidth is more than enough.

All nodes, except sources and destination nodes are defined as MPLS node. The set of node {Src_0, Src_1, Src_2, Src_3, Src_4, Dst_18, Dst_19, Dst_20, Dst_21} are all defined as standard node in NS-2, and the remaining nodes are MPLS capable node.

3.2 Traffic Model

The flows of traffic VoIP is defined as Constant Bit Rate (CBR). CBR is a type of application layer traffic in which the data rate sent by traffic source is constant. CBR is widely used by a service that requires predictable response time and CBR is run on the link with static bandwidth definition. The example of CBR traffic is video conference traffic, VoIP traffic, and traffic generated from on demand service.

In the simulation, flows of VoIP traffic and Best Effort are run on top of UDP transport protocol. However, to carry data traffic, the flow is run on top of TCP. The mapping of each traffic flow onto Bandwidth Constraint Model is defined as follows:

- VoIP traffic will be mapped onto a flow with Flow ID. This Flow ID 1 I then mapped onto Class Type 1 (CT1) to allow VoIP traffic to get the portion of BC1.
- Data traffic will be mapped onto a flow with Flow ID 2. This Flow ID 1 is mapped onto CT2 with BC2 allocation model.
- The remaining Best Effort Traffic will be mapped onto Flow ID 3, which is then allocated for CT3. Based on simulation model, CT 3 is allocated using BC3 model

4. PERFORMANCE EVALUATION

The goal of [16] is to identify and develop simulation requirement to evaluate bandwidth constraint model in DS-TE environment. This paper continues [16] by running the simulation based on several scenario. The simulation in this paper is to find the performance overview and analysis the effect of bandwidth constraint management to the performance metrics, such as throughput, end to end delay, packet loss and bandwidth utilization in DS-TE network. We define two scenario for simulation run, i.e. LSP is governed automatically before traffic exist in the DS-TE network, and the second one is traffic flows before LSP is governed in DS-TE network.

Throughput X (bps) is obtained from the calculation of succeeded traffic sent of every class of traffic to the destination nodes (AcceptedTr(dst)). The obtained value will be divided by the length of *n* simulation time. In NS-2 simulator, we define throughput by

$$X = \frac{\sum AcceptedTr(dst)}{n} \quad (1)$$

Packet Loss (PLossTR) is the number of dropped packet during the length of n simulation time for every flow of traffic. In NS-2 simulator we define the number of packet loss by

$$PLossTr = \frac{\sum DroppedFlow}{\sum SentFlow} \quad (2)$$

In this simulation, we define end to end delay as the time needed for every transmitted packet to reach the destination. The delay component that forms the end to end delay is transmission delay and queuing delay.

The result of simulation of scenario 1, which is the throughput of application, is depicted in Table 1a and 1b. In both scenarios, we can see that the effect of Bandwidth Constraint Model using MAM can improve the throughput of network application, significantly for VoIP and Data traffic.

Table 1 Average Throughput in Scenario 1

	Voice	Data	Best Effort
DSTE MAM	1275221	39487	15833
NON DSTE	1157055	1687	15833

Table 2 Average Throughput in Scenario 2

	Voice	Data	Best Effort
DSTE MAM	1700485	59647	47500
NON DSTE	1550060	3367	833

However, in scenario 2 we have more packet loss because the traffic flows in the network before LSP is governed. This means that in that condition, traffic flows without using path definition (LSP) and the routing path used to forward the traffic is only using OSPF as the default routing protocol.

Figure 4. Bandwidth Utilization

Figure 4 shows bandwidth utilization as the result of simulation, which is measured from node 5 to node 17 in the simulated topology. During simulation time 0 – 1.1s, bandwidth utilization takes only 4% on both DS-TE and Non DS-TE network. This is because in this phase, the available bandwidth is only utilized for flooding of routing packet. After 3s, DS-TE shows better performance in term of utilizing the available bandwidth by defining new path instead of the existing path governed previously using OSPF. Hence the bandwidth utilization on this path is up to 78% at simulation time 5s. However, for Non DS-TE, the path used to forward the IP packets is still based on OSPF, and the bandwidth utilization can only reach 43% at simulation time 5s.

5. CONCLUSION

Based on two scenarios of simulation, it can be concluded that Maximum Allocation Model (MAM) as the Bandwidth Constraint Model in DS-TE affect better performance compared to the original MPLS-TE. During the simulation, MAM can increase throughput up to 10%, reduce packet loss up to 10%, and reduce the end-to-end delay significantly for VoIP application, i.e. up to 38%. The effect of MAM also influenced the bandwidth utilization.

From the simulation, the creation of LSP in MAM is also the key factor in DS-TE. MAM is possible to be used to govern the LSP automatically per class basis. Thus, every defined LSP can be used to stream one class of traffic flow to significantly increase the throughput.

6. REFERENCES

[1] Alvarez, Santiago. QoS for IP/MPLS Networks. Cisco Press, IN, 2006

[2] Awduche, D., et al.: Requirements for Traffic Engineering over MPLS. IETF RFC 2702, September 1999

[3] F. Le Faucher et al., Requirements for Support of Differentiated Services-aware MPLS Traffic Engineering, IETF RFC 3564, July 2003

[4] F. Le Faucher, Maximum Allocation Bandwidth Constraints Model for Diffserv-aware MPLS Traffic Engineering, IETF RFC 4125, June 2005

[5] Blake et al , An Architecture for Differentiated Services, RFC 2475, December 1998

[6] F. Le Faucher, Protocol Extensions for Support of Diffserv-aware MPLS Traffic Engineering, IETF RFC 4124, June 2005

[7] E. Rosen et al., Multiprotocol Label Switching Architecture, IETF RFC 3031, January 2001

[8] http://en.wikipedia.org/wiki/RSVP-TE accessed on June, 20 2009

[9] Fall, Kevin dan Varadhan , Kannan I. The NS Manual. .2009

[10] E. Rosen et al., Multiprotocol Label Switching Architecture, IETF RFC 3031, January 2001

[11] Park, Kun I. QoS in Packet Networks. Springer, US, 2005

[12] Osborne, Eric dan Simha, Ajay. Traffic Engineering with MPLS. Cisco Press, IN, 2002

[13] http://www.isi.edu/nsnam/ns/doc/index.html

[14] Davide Adami, Christian Callegari, Stefano Giordano, Fabio Mustacchio, Michele Pagano, Fabio Vitucci "Signalling Protocols in DiffServ-aware MPLS Networks: Design and Implementation of RSVP-TE Network Simulator" IEEE Global Telecommunications Conference (GLOBECOM 2005), Nov 28 - Dec 2, St. Louis, MO, USA

[15] Davide Adami, Christian Callegari, Stefano Giordano, Michele Pagano "A New Path Computation Algorithm and its Implementation in NS2" IEEE

[16] Deni Sartika,Bayu Erfianto,Tribroto Harsono.2010. Desain Simulasi Bandwidth Constraint Model Pada Diffserv-Aware MPLS TE. Tugas Akir Fakultas Informatika, Institut Teknologi TELKOM 2010

User Accounting System of Centralized Computer Networks Using RADIUS Protocol

Heru Nurwarsito
Electrical Engineering Department, Faculty of Engineering, UB.
Jl. MT. Haryono 167 Malang
Indoenesia 65145
herunur@gmail.com

Raden Arief Setyawan
Electrical Engineering Department, Faculty of Engineering, UB.
Jl. MT. Haryono 167 Malang
Indoenesia 65145
rariefset@ub.ac.id

Handoko D Fatikno
Electrical Engineering Department, Faculty of Engineering, UB.
Jl. MT. Haryono 167 Malang
Indoenesia 65145
koko02_cuakep@yahoo.co.id

ABSTRACT
There are two types of connections on the local computer network (Local Area Network [LAN]) is through a cable (wired) and wireless (wireless). But we don't know with certainty who the user is connecting to the network. This can threaten network security for users who do not connect the carrying and not easy to trace and find out who did it if there is an any attack. Based on security, we need a security system Authentication, Authorization, and Accounting (AAA) to ensure the safety and convenience of users, because with this system we know with certainty who the user is connecting. The user can move from one network to another network, it needed a centralized system. Therefore, in this research using RADIUS protocol and focus on the accounting process. Accounting results in this study were analyzed and can be used for tracking purposes (who, when and where the request originated) if there is a crime that uses an internal computer network access of University of Brawijaya, the process of user access fees if you use Internet access, as well as materials for planning and resource allocation of bandwidth at the time mendatang.Perancangan software includes logical design of computer networks, using a RADIUS server freeradius design, design of database server for storing accounting data, the design of the NAS as a RADIUS client. With this system can be produced with enhanced security requirements in the form of tracking IP addresses and time of occurrence is known, the value decreased throughput after applying this system that is equal to 5.75%, we can utilize the results of accounting data for purposes of calculating the cost of user access to computer networks.

Keywords
Internet Network Security, AAA, RADIUS, accounting, centralized, Protocol

1. INTRODUCTION
These last few years, information technology and telecommunications in Indonesia began to develop rapidly. The more sophisticated and modern and affordable device to Indonesia such as notebooks, PDAs, and cell phone equipped with the facilities to be connected to a computer network can be a factor. Some of the activities of life such as business, buying and selling goods, education, exchange information, news, daily records, banking transactions, communicate with other people far away place, all developing in the direction the Internet.

Problems to be faced when connected to computer networks and the Internet is all about safety. Today many systems found in computer networks that do not implement an adequate security system so it is possible for users who are not entitled to (illegal) can enter into the computer network system is connected. The intruder may have committed acts detrimental to retrieve data such as, attacking other computers or devices connected to that network and others. With the growing need for information and telecommunications, we need a reliable network security method and can monitor user activity connected to the computer network. For it's one way we can use the AAA security mechanism which is a continuation of the Authentication, Authorization and Accounting [[5]: 5].

Authentication is the activity of a person to verify that something is valid or invalid [[5]: 5]. Authorization is the determination of whether something (a user or device) that has the license for accessing the service. [[5]], and tracking of significant accounting of who, what, when and where the request originated and where the response will be sent [[5]: 5]. AAA is very important in wireless networks because of the roaming and user identity to be kept, and the system must continue to monitor user activity [[5]: 5].

Remote Access Dial-In User Service (RADIUS) is one of the AAA network protocols in use enough to dominate in the field. This is because the protocol is open and vendor independent [[4]: 6]. There are several reasons why RADIUS is selected, that is simple, efficient, and easy to implement [[3]: 3]. RADIUS runs a centralized user administration system, this system will simplify the task administrator. With this system users can use the hotspots in different places by performing authentication to a RADIUS server [[3]: 3].

Base on the concept of RADIUS and accounting processes going on inside as well as designing the implementation of user accounting system is centralized computer network using RADIUS protocol at University of Brawijaya (UB) and analyzing accounting processes that occur in it. Therefore in this study using RADIUS protocol and focus on the accounting process because for the authentication and authorization process will be discussed in another study done in conjunction with this research by Winda Septarini sisters. Accounting results in this study were analyzed and can be used for tracking purposes (who, when and where the request originated) if there is a crime that uses an internal computer network access UB, the process of user access fees if you use the Internet access.

2. MECHANISM OF AAA

Mechanism of AAA (Authentication, Authorization, Accounting) mechanisms regulate the procedure how to communicate, both between the client to the network domains as well as between clients with different domains while maintaining the security of data exchange. [[1]: 21]

2.1 Authentication (Authentication)

Authentication is the process of ratification of the identity of the customer (end user) to access the network. This process begins with sending a unique code (eg, username, password, pin, fingerprints, etc.) by the subscribers to the server. On the server side, the system will receive a unique code, then compare it with a unique code stored in the database server. If the result is the same, then the server will provide access rights to the customer. But if the results are not equal, then the server sends a failure message and refuse customers access rights [[4]: 4].

2.2 Authorization (Authorization)

Authorization (Authorization) represents the process of checking the authority of the user, anywhere access rights allowed and which are not [[1]: 22].

2.3 Listing (Accounting)

In this case, the Company recorded a network resource that has been used by the customer. This recording can be aimed to the analysis, inspection, payments, planning and resource allocation. Recording can also be to secure systems, such as watching for suspicious customers, and so forth [[4]: 5].

Listing (Accounting) is the process of data collection information about how long the user to connect and billing time has passed during usage. [[1]: 22]

3. REMOTE AUTHENTICATION DIAL-IN USER SERVICE (RADIUS)

RADIUS is a protocol developed for the process of AAA (authentication, authorization, and accounting). RADIUS runs a centralized user administration system, this system will facilitate the task of administrators [SET, 05: 3].

Remote Access Dial In User Service (RADIUS) developed in 1990 by Livingstone dipertengahan Enterprise (now Lucent Technologies). At first the development of RADIUS use port 1645 which turned out to have clashed with datametrics service. Now the port that is used RADIUS is port 1812 which is default format on the Request for ditetapakan Command (RFC) 2138 [[1]: 22].

Remote Authentication Dial-In User Service (RADIUS) (RFC2865) was originally used by ISPs to otenti? Cation with the username / password before you can connect with the ISP network, and it usually is for dial-up users. RADIUS is used to back-end server in the 802.1x authentication. In addition there are several protocols RADIUS AAA (Authentication, Authorization, Accounting), which can also use the TACACS, TACACS +, and DIAMETER [FUA-08: 6].

3.1 RADIUS Data Packet Format

RADIUS packet format consists of the Code, Identifier, Length, authenticator and Attributes as shown in the following figure [[1]: 22]

Figure 1. RADIUS data packet structure

.

Principle Radius

RADIUS is a security protocol that works using a distributed client-server system that is widely used with AAA to secure the network users who are not entitled to [[1]: 22].

RADIUS to authenticate users through a series of communications between clients and servers. If a user has successfully authenticating, the user can use the services provided by the network [[1]: 22].

RADIUS protocol is a UDP-based connectionless protocol that does not use a direct connection. One RADIUS packet is marked with a field that uses UDP port 1812. Some considerations RADIUS uses UDP transport layer, namely: [[1]: 25]

* If the first authentication request fails, then the second request should be considered.
* Nature stateless protocol that simplifies the use of UDP.
* UDP simplifies the implementation of the server side.

3.2 RADIUS Access Mechanism

Traffic protocol messages on the RADIUS request and response methods (client / server) that can be seen in the image below [[1]: 27].

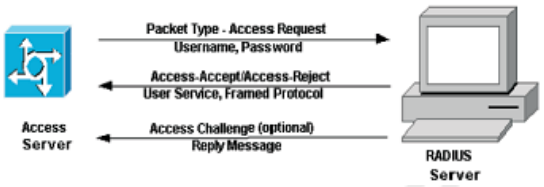


Figure 2. Traffic messaging client NAS with RADIUS server

* Users do dial-in on the NAS. NAS will prompt the user to enter name and password if the connection is successfully established. [[1]: 25].
* NAS will be building a data packet of information, called the access request. Information provided by the NAS in the RADIUS server contains specific information from the NAS itself that requested the access request, the port that is used for modem connections as well as names and passwords. For protection from hackers, the NAS which acts as a RADIUS client, encrypts the password before being submitted to the RADIUS server. Access requests are sent on the network from the RADIUS client to the RADIUS server. If the RADIUS server can not be reached, moving the RADIUS client can route on an alternative server if the configuration is defined in the NAS [[1]: 25].
* When the access request is received, the authentication server will validate the request and decrypt the data packet to obtain information on names and passwords. If your name and password as the database server, the server sends access accept that contain information network system needs to be provided by the user. Also accept this access can contain information to restrict access to users on the

network. If the login process does not meet the fitness, then the RADIUS server sends the access rejection at the NAS and the user can not access the network [[6]: 26].

To ensure a user request actually given on the right side, the server sends RADIUS authentication key or a signature that indicates its presence on the RADIUS client [[1]: 26].

3.3 Radius Architecture

This system architecture consists of three parts, namely a wireless node (Supplicant), access point (autentikator), authentication server. Authentication server used is the Remote Authentication Dial-In Service (RADIUS) server and is used to authenticate users who will access the wireless LAN. EAP is a layer 2 protocol that replaced the PAP and CHAP [[3]: 5].

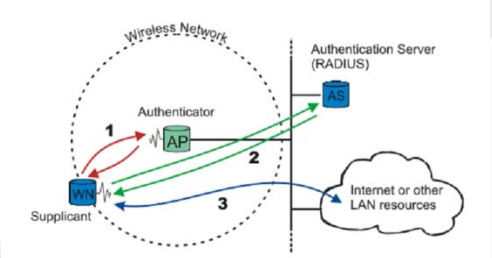


Figure 3. Architecture using a RADIUS server Authentication Mechanism

Explanation of Figure:

* Wireless Node (WN) / Supplicant to request access to the wireless network, Access Points (APs) will be asked Supplicant identity. None other than the data traffic is allowed before the Supplicant EAP terautentikasi. Access Point is not a autentikator, but the access point contains autentikator [[3]: 5].
* Once a user name and password is sent, the authentication process begins. A protocol used between the Supplicant and Autentikator is EAP, or EAP over LAN protocol (EAPoL). Autentikator encapsulates the EAP message back into the RADIUS format, and sends it to the RADIUS server. During the authentication process, autentikator only convey packets between Supplicant and the RADIUS server. After the authentication process is completed, RADIUS server sends a message of success (or failure, if the process fails autentikasinya) [[3]: 6].
* If the authentication process successfully, Supplicant allowed to access the wireless LAN and / or the Internet [[3]: 6].

3.4 RADIUS Accounting

Basic document standard RADIUS (RADIUS2865) does not provide specifications for accounting support. Akantetapi, RADIUS accounting RFC information is defined in another section (RADACC2866),. Accounting procedures are also based on client-server model where the client (NAS) pass information about the user to the RADIUS accounting server, which is the host machine from the RADIUS accounting [[2]: 139].

RADIUS Accounting is using two types of messages: Request and Accounting Accounting Response, both are also sent via UDP. Accounting Request is always sent from the RADIUS client to the RADIUS server, when the Accounting Response generated by the RADIUS server when receiving and processing the Accounting Requests (Figure 7.4.3). However, as we shall see in the roaming scenario of proxy server may play a role in the exchange of Request Accounting-Response [[2]: 139].

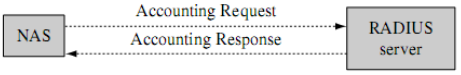


Figure 4. Traffic messaging client NAS with RADIUS server

3.5 Accounting Mechanism

NAS is able to support RADIUS accounting mechanism to generate Accounting Request "Start" at the time of operation start and send it to the RADIUS accounting server. This package is to determine, among other things, the type of service delivered, and the user that sent the service. Upon receipt of a valid accounting request, the server adds accounting records in the log and respond to requests generated Accounting Response to indicate that the packet was received [[2]: 139].

At the end of service delivery, the client will generate Accounting Stop packet describing the type of service has been delivered and such statistics and the duration of the session, the reason for disconnect, the number of inputs and output.itu octets will all be sent to the RADIUS accounting server, which will send a return statement that the packet has been received. Of course, if the RADIUS accounting server can not successfully record, accounting package will not send notice of Accounting-Response to the client [[2]: 140].

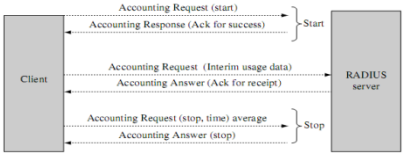


Figure 5. Exchange of Messages at the time of accounting session

4. DESIGN

In this chapter will explain the design of implementation mechanisms accounting system is centralized computer network users using the RADIUS protocol UB. Before designing the writer must know beforehand how the existing computer network at UB today. Next, we conducted a needs analysis and system design.

4.1 Existing Computer Network UB

This section will explain how the topology of computer networks such as UB and what specification the routers that exist on computer network systems UB today. This is needed for the analysis of where the laying of NAS and servers as well as the extent to which this system can be implemented at UB will relate to the specification of hardware and software that already exists. [8]

At present, there are two types of connections on the local computer network (Local Area Network [LAN]) UB is through a cable (wired) and wireless (wireless). For computers

that use cable networks are the main network (backbone) in each faculty and from each faculty were divided again into every major. After that there are majors at some point access point hotspots is provided as an area for students who want to connect to computer networks wirelessly.
UB Computer Network Topology

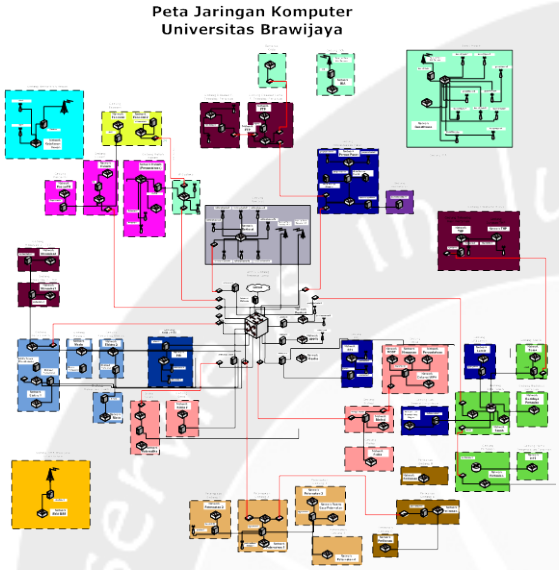


Figure 6. Computer Network Topology UB

From the above figure we can know that the type of computer network topology at UB is a type of tree networks. It can be seen from the structure of the network, which combines characteristics of linear networks and star networks. Computer network at UB consists of a set of workstations berkonfigurasi with star structure connected to the backbone network which is the main network. Backbone network or the major networks are star-shaped UB where his client node is a gateway from a network computer in every faculty and non-faculty. While every major gateway connected to a linearly connected to the network backbone, which was on the faculty. Furthermore, for networks in the majors using a star network is connected to the gateway majors.

4.2 Specification network routers UB

Routers at UB still use a router with a linux pc as the operating system. Some still use the Pentium III uses even one who is still using the Pentium II, while for others already using the Pentium IV and above. Referring to the specifications of the hardware requirement analysis in Section 4.1.1.1 it is almost 50% of the routers that exist at UB has fulfilled the requirement.

4.3 UB Computer Network Users

Basically, users are allowed to use the facilities computer network is the UB faculty, staff, and students of UB. But, along with the increasing number of access points that provide free wireless Internet access at this UB.saat, we can not know with certainty who the user is connecting to a computer network UB. This is because each person can freely connect to computer networks UB through the access point without security systems are adequate, so it is possible for users who are not entitled to (illegal) can enter into the computer network system linking and do an activity that is harmful.

4.4 Needs Analysis

After seeing the UB existing computer network can be defined several requirements as follows:

* A system that can perform user registration process anyone who is connected
* A system that can do recording of user activity connected.
* The system is easily implemented with existing computer networks UB

4.5 System Design

The Design Implementation of Computer Network User Accounting System By Using Centralized RADIUS protocol of this study are based on analysis of needs that have been discussed above. Pearancangan include how the topology of the system and working principle of the system will then for the purposes of testing the system, then created a prototype of a system which was then implemented to analyze how the performance of these systems have been designed.

4.6 UB Networks Designing Accounting System

UB network accounting system designed centrally, ie all the NAS or the radius of an existing client-router at each end will make the process of authentication, authorization and accounting on a RADIUS server. It aims to each user can use the facilities UB computer network resource usage wherever he was in the campus area.

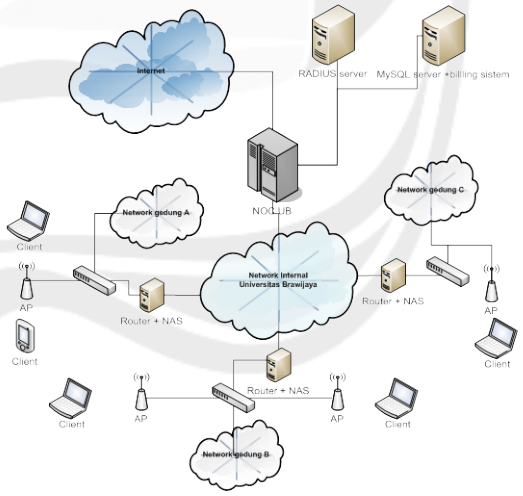


Figure 7. Designing Computer Network Topology UB accounting system

RADIUS servers, database servers, and billing system is placed on the NOCs (Network Operation Center) - UB which is a computer network operation center at UB. Next to each end-existing router at UB are used as a NAS that serves as the gateway and the user's computer networks UB autentikator well-connected using a cable or wirelessly. While for the access point will function only on layer 2 alone, it aims to be easy in the future because it does not restrict the implementation of the brand and certain types of access points.

For more details we can see in the picture the following topology.

4.7 Prototype Design of Accounting System Computer Network UB

Designing a network accounting system prototype UB kumputer made based on the analysis previously discussed. The design of this prototype will be used for testing systems that have been designed. The result of the prototype design will then be implemented in a computer network Electrical Engineering Department of UB to the author can obtain real data for analysis of test results. In this prototype design topology used is the form of a tree network, this computer network topology disesuakan with UB. In this prototype laying RADIUS server, MySQL server and billing system is adapted to the design that is made before the NOC-UB zone. Subsequently used by the two end-routers that function as NAS for testing two types of computer networks and the cable network is a network without cables. More detail can be seen in the following figure.

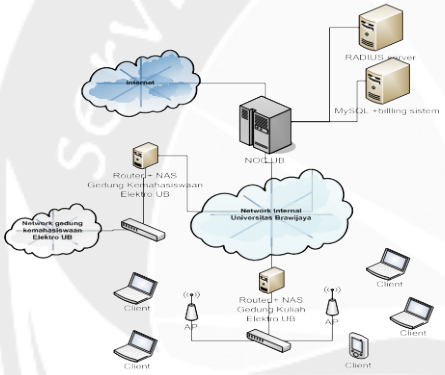


Figure 8. Prototype Design of Accounting System Computer Network UB

The allocation of IP can be seen in the following table:

Table 1. IP Allocation Prototype System for Computer Network UB accounting

Name	*Interface*	IP Address	Gateway
Server RADIUS	eth 0	172.18.3.3 1/24	172.18.3.1/24
Server basis data	eth 0	172.18.3.3 2/24	172.18.3.1/24
NAS-1	eth 0 (*uplink*)	172.17.67. 62/	172.17.67.41/
	tun 0 (*downlink*)	172.17..91. 1/25	
NAS-2	eth 0 (*uplink*)	172.17.67. 67/	172.17.67.65/
	tun 0 (*downlink*)	172.17.91. 0/25	

4.8 The Design Specification Hardware

Hardware specification is a description of the device or equipment used in this study. Specification contains the processor, memory, hard disks and other equipment. Description of the device could also mention the brand, type, name, speed, capacity and other matters that sufficient detail of the device.

The hardware used in this study is the fruit of 4 personal computer used as a RADIUS server, 2 RADIUS client and server database that will be developed as a prototype computer network at UB for the purposes of testing. In addition, the required 2 wireless access points for testing in order Supplicant may be connected to a computer network using either cables or wireless.

4.9 The Design Specifications Software

At the user's use RADIUS accounting system is the software used that is:

Freeradius

Freeradius is an application server that is used as a RADIUS server authentication, authorization and accounting following its design specifications

* Freeradius using version 2.1.6-0
* What is the center of UB internal network.
* Server has support for MySQL as a database server that stores the user's accounting data.
* Server can restrict users so that can not connect with the same account at once.
* The server can perform user access restrictions based on duration or volume of usage.
* Freeradius register the IP address of each of the existing NAS to communicate with the server. For the security of the connection between NAS and server, both having a similar keyword called radiussecret.

Mysql

MySQL is a DBMS (Data Base Management System) that serves as a database server for storage of accounting data from RADIUS server. The following design specifications

* MySQL version 5.0
* A separate database server with RADIUS servers and are one segment of the IP address to be able to communicate with each other.
* What is the center of UB internal network.
* The database has a table easyhotspot scheme which is a modified scheme Radius table with additional tables for user management.

Chillispot

Chilispot is an application that is used as a NAS which is a RADIUS client. The following design specifications

* Use version-1.1.0-1 Chillispot
* NAS IP address registered with the RADIUS server.
* Functioning as a router and DHCP server can assign IP addresses automatically to the user.
* Having an IP address and network segmentation similar to the existing network UB.
* Firewall to use the existing rule with additional rules UB network that supports the functions of NAS is to ridirect all packet forwarding to port 3990 before users authenticate.
* To secure data communications between the NAS and the user then applies the function https login page, so that valuable information such as usernames and passwords in an encrypted state.

Design of User Management

In accordance with the above requirement analysis, the management of computer network users are divided as follows UB

* Students
  - Username to use NIM
  - Connection will drop out if the user does not do an activity for 5 minutes
  - Users can not login more than once at the same time.
* Lecturer and Staff
  - Username used Nip / Nik
  - Connection will drop out if the user does not do an activity for 5 minutes
  - Users can not login more than once at the same time.

5. IMPLEMENTATION

The prototype system is a computer network user accounting system that has been designed UB in Chapter IV, will be implemented and tested in this phase. There are three major steps undertaken in this chapter, namely implementation, testing and analyzing the results of user accounting system prototype UB computer network.

Configuration Implementation is done by doing the installation of hardware and software perangakat. After that, prepare in accordance with the prototype of a computer network topology that has been made in the design stage. Configuration is then performed on each server as follows.

5.1 RADIUS Server

Installing the RADIUS server configuration on the computer after the Linux Operating System is installed Ubuntu. In the installation, use the minimum mode, ie, only install the required software packages and, for reasons of security and brevity. Selected in the installation include:

* Freeradius, RADIUS server application.
* Freeradius-mysql, freeradius additional modules that can use MySQL as database server.

Configuring the RADIUS server needs to be done in order to function properly is as follows

* Configure the default gateway IP address and RADIUS server
* Conduct freeradius configuration in order to communicate with the MySQL database server
* Conduct freeradius configuration in order to read and store data from the user accounting database server
* Conduct freeradius configuration to restrict the user, so can not connect with the same account at the same time
* Conduct freeradius configuration to be able to perform user access restrictions based on duration or volume usage
* Collecting information on the NAS are allowed to communicate with RADIUS servers

7. CONCLUSIONS

Base on the testing results and design analysis implementation of accounting systems in a centralized computer network users using the RADIUS protocol, it can be concluded that:

- Processes that occur on a network accounting system prototype UB computer is consistent with existing theory on the basis of the theory.
- Obtained completely user accounting data including accounting data numbering plan data, user session id, unique user id, username to connect, nas port used by a user id, nas port type used by the user, the start time of the user to connect, end time users to connect, long time users to connect, the user input data used, the amount of output data that is used the user, nas mac address of the gateway user, mac address of the user who made the connection, user connection because he decided, and when a user ip address dataconnections.
- We can produce security enhancements provided in the form of tracking IP addresses and time of occurrence unknown.
- The value is applied after the system throughput decreased by 5.75% compared with the value before the application system throughput.
- We can utilize the results of accounting data for purposes of calculating the cost of user access to computer networks.

8. REFERENCES

[1] Arif, Teuku Yuliar; Syahrial; and Zulkiram. 2007. Study on Authentication Protocol Service Internet Service Provider (ISP). Journal Elektrika Engineering: Electrical Engineering Department - Faculty of Engineering, University of Syiah Kuala.

[2] Nakhjiri, Madjid and Nakhjiri, Mahsa. 2005. AAA And Network Security for Mobile Access (RADIUS, DIAMETER, EAP, PKI and IP Mobility. John Willey & Sons Ltd.: England.

[3] Setiawan, Agung W. 2005. Remote Authentication Dial In User Service (RADIUS) Authentication for Wireless LAN Users. Faculty of Electrical Engineering Department of Industrial Technology of Bandung Institute of Technology.

[4] Warsito. 2004. Network Security System Using Multi-Domain Protocols DIAMETER. Master Program in Electrical Engineering Special Programs Information Technology - Institut Dikmenjur

[5] Wicaksono, Sulistyo Excellence. 2006. Study of Network Security Systems CDMA. Faculty of Electrical Engineering, Institut Teknologi Bandung

[6] Kizza, Joseph Migga. 2005. Computer Network Security. Springer: United States of America.

[7] pradhana, Harindra Vishnu. 2006. Web Design Basics with PHP and MySQL. Electrical Engineering, University of Diponegoro.

[8] Supriyadi, Andi; and Gartina, Dhani. 2007. Choosing a Network Topology and Hardware in Design A Computer Network. Journal of Agricultural Informatics Volume 16 No. 2, 2007.

Wireless Data Communication with Frequency Hopping Spread Spectrum (FHSS) Technique

Khin Swe Myint
Computer University (Banmaw)
No.48-41st Street, Botataung Township
Yangon, Myanmar
095-09-99-25198
papanono@gmail.com

Zarli Cho
Computer University(Pyay)
24-B(3rd floor),Nanattaw Road
Kamayut Township,Yangon, Myanmar
095-09-22-00589
chitsu2010.2@gmail.com

ABSTRACT
This paper attempts to achieve a wireless data communication between two or more personal computers (PCs) by using the Frequency Hopping Spread Spectrum (FHSS) technique. A circuit containing of a PIC16LF877A microcontroller and an nRF905 single-chip radio transceiver is used as a transceiver unit for controlling and handling frequency and data. The PC uses the RS-232 interface to communicate with the transceiver unit. The PIC16LF877A at the transceiver unit uses the Serial Peripheral Interface (SPI) protocol to interface with the nRF905. The PIC controls the hopping frequency and timing of the transceiver unit. It uses the predefined pseudo-random number algorithm to change the frequency. The nRF905 transmits/receives data with the hopping frequency set by the PIC with Manchester-encoded Gaussian Frequency Shift Keying (GFSK) modulation. The master transceiver unit generates the time synchronization pattern for the slave transceiver unit and the slave transceiver unit gets the time synchronization pattern from the master transceiver unit to hop frequency simultaneously with the master.

Keywords
Pseudo random number, FHSS, SPI

1. INTRODUCTION

Today, communication has increasing influence on our daily life. Wireless data communication services allow people to access the data network without a physical connection. Wireless communication or communication without the need for physical contact between sender and receiver is a small scaled data communication system utilizing radio frequencies rather than cable.

Communication by radio means the transfers of intelligent from one point to another through space using radiated electromagnetic wave in the frequency spectrum of from about 10kHz to 300GHz. Electromagnetic waves (radio waves) are propagated through space from the transmitting antenna to receiving antenna. Radio frequency (RF) technology is implemented using two main methods. They are narrow band and spread spectrum (SS) techniques. Since narrow band modulation schemes have problem with multipath transmission and they are very sensitive to interference, SS technology is preferred. The more resent frequency is the spread spectrum. SS communications systems are often used there is a need for message security and confidentiality.

2. SPREAD SPECTRUM TECHNOLOGY

In modern telecommunications spread spectrum methods are widely present. The term spread spectrum has been used in a wide variety of military and commercial communication systems. Spread spectrum communication is mostly digital communications technology. Spread spectrum radio communication, long a favorite technology of the military because it resists jamming and is hard for an enemy to intercept, is now on the verge of potentially explosive commercial development. Because spread spectrum signals which are distributed over a wide range of frequencies and then collected onto their original frequency at the receiver, are so inconspicuous as to be 'transparent'. They are unlikely to be intercepted by military opponents. Spread spectrum uses wide band, noise-like signals. Because spread spectrum signals are noise-like, they are hard to detect. Spread spectrum signals are also hard to intercept or demodulate. Further, spread spectrum signals are hard to jam (interfere with) than narrow band signals. Spread signals are intentionally made to be much wider band than the information they are carrying to make them more noise- like. Spread spectrum signals use fast codes that run many times the information bandwidth or data rate. These special 'Spreading' codes are called "Pseudo Random" or "Pseudo Noise" codes. They are called "Pseudo" because they are not real Gaussian Noise. The use of these special pseudo noise codes in spread spectrum communications makes signals appear wide band and noise-like. At the receiver the signal is 'despread' using a synchronized replica of the pseudo-noise code. It is this very characteristic that makes SS signals possess the quality of Low Probability of Intercept. General model of spread spectrum digital communication system are shown in Figure 1.

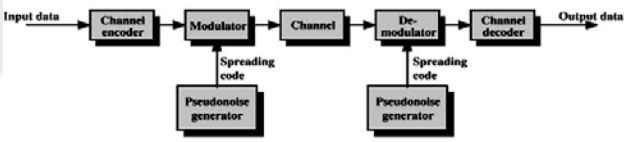


Figure 1. General Model of Spread Spectrum Digital Communication System

Spread spectrum systems have been classified by their architecture and modulation. The commonly employed SS modulation techniques are the following.

- Direct Sequence Spread Spectrum (DSSS)
- frequency hopping spread spectrum (FHSS);
- hybrid direct-sequence and frequency-hopping spread spectrum;
- time hopping spread spectrum and
- chirp spread spectrum

IEEE 802.11 is being proposed to provide wireless connection for local area network using spread spectrum techniques in the Industrial, Scientific, Medical (ISM) bands. In this paper build wireless communication system operating at 867.2-870.2 MHz of the ISM bands. The physical layer of this wireless data communication has designed to perform frequency hopping spread spectrum processing and RF transceiving.

3. FREQUENCY HOPPING TECHNIQUE

Frequency hopping is not a signal spread across the spectrum, but a broad bandwidth in the spectrum which is divided into many possible broad cast frequencies to which the data will be sent over. The wide band frequency spectrum desired is generated in a different manner in a frequency hopping system. That is it "hops" from frequency to frequency over a wide band. For FHSS, there exists a code which determines at any particular moment in time what frequency it will transmit at hopping from frequency to frequency. FHSS systems transmit data on one frequency for a period of time, before hopping to another frequency to continue the transmission. Hence the only way to obtain the transmission is to be has an identical code that knows which frequency it will jump to next. The frequencies are used in a predefined pseudo-random sequence that both the master and slave know. The hopping pattern or sequence appears random but is actually a periodic sequence tracked by master and slave. The hopping pattern of the frequency synthesizer is determined by the output of the PN generator. Random frequency hopping sequence is shown in Figure 2.

Thus, a FHSS system produces a spreading effect by pseudo-randomly hopping the RF carrier frequency in the available RF band. The processing gain of the FHSS system is given by the ratio of the spread signal bandwidth to the information bandwidth.

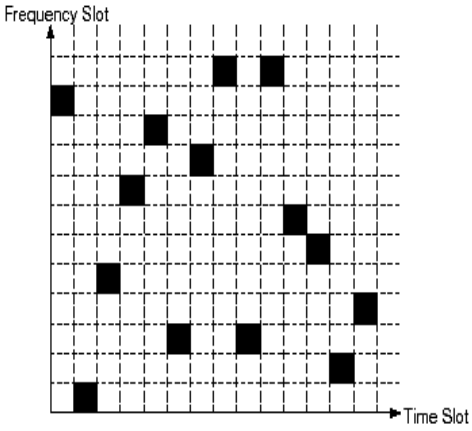
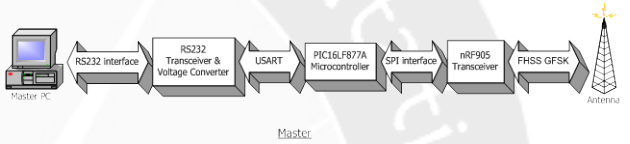
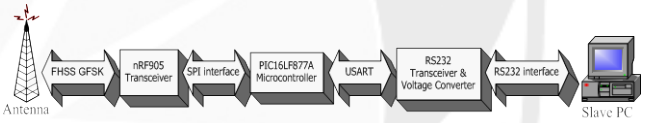


Figure 2. Frequency Hopping in the Time and Frequency Domains

4. SYSTEM DESIGN



(a) Master Section



(b) Slave Section

Figure 3. Block Diagram of the Wireless Data Communication with FHSS Technique

The goal of this system is to find out how data could be transferred wirelessly from primary station to remote stations link physically separate computers. The entire block diagram of the system is shown in Figure 3.

At the master, the PC generates the data. Which uses the RS-232 interface to communicate with the transceiver unit. USART is used to helps PIC link to PC. The PIC 16LF877A at the transceiver unit uses the Serial Peripheral Interface (SPI) protocol to interface with the nRF905. PIC 16LF877A microcontroller and nRF905 radio transceiver are used as a transceiver unit for controlling and handling frequency and data. The nRF905 single chip transceiver transmits the signal to the antenna by using GFSK modulation technique.

At the slave, the incoming signal is picked up by the antenna and this section used the same technique as the master, but in reverse. This slave transceiver unit gets the time synchronize from the master unit to hop frequency simultaneously with the master. It uses the predefined pseudo-random number algorithm to change the frequency. This system uses large number of frequencies 867.2-870.2 MHz ISM bands for the RF transmission. It can switch 16

hopping channels and each channel is 200 kHz wide and the data rate is 50kbps.

5. HARDWARE AND SOFTWARE IMPLEMENTATION FOR FHSS SYSTEM

5.1 Hardware Implementation

It is not easy to handle high frequency RF components without using high end equipments and fabrication tools. Every millimeter length of copper track can effectively change the signal strength. Because this system emphasize on the transferring data with carrier frequency hopping technique, precise carrier frequency generation is needed and which can be changed frequently. RC oscillator carrier frequency generators cannot be used because their precision is not satisfactory in rapid changing mode. Thus, crystal based frequency generation is used in this design. nRF905 multiband RF transceiver from Nordic Semiconductor, U4, is chosen as RF transceiver unit. It is a single-chip radio transceiver for the 433/868/915 MHz ISM band which runs on crystal based clock.

nRF905 working voltage is between 1.9V and 3.6V. To simplify the hardware design, 3V power supply is decided to be used for the whole system. Thus all components must be worked on 3V power supply. A PIC16LF877A microcontroller from Microchip Corporation, U3, is used as the main controller of the system because LF series of PIC microcontrollers operate on 2V to 5.5V range. MAX-232ACPE RS-232 transceiver from Maxim Integrated Product, U1, is used to interface between the microcontroller and the PC (personal computer). Although the nominal working voltage of U1 is 5V, it was found working well with 3V power supply.

5.1.1 nRF905

The nRF905 transceiver consists of a fully integrated frequency synthesizer, receiver chain with demodulator, a power amplifier, a crystal oscillator and a modulator. The ShockBurst feature automatically handles preamble and CRC. Configuration is programmable by use of the SPI interface. Current consumption is very low, in transmit only 11mA at an output power of -10dBm, and in receive mode 12.5mA. In this design, nRF905 is designed to work with carrier frequency between 867.2MHz to 870.2MHz in 200kHz step channels.

A 20MHz crystal, Y2, is used as a working clock. Resistor R_6, capacitor C_{21} and C_{22} are crystal oscillator bias components. Capacitor C_9, C_{11}, C_{12} and C_{16} are power supply decoupling capacitors. R_4 is a reference bias resistor. Capacitor C_{18} and C_{20} are PA supply decoupling capacitors. Resistor R_5 is used for antenna Q reduction. Capacitor C_{15}, C_{17} and C_{19} are antenna tuning capacitor. All resistors and capacitors mentioned above are recommended by Nordic for 868MHz carrier frequency range.

The ANT1 and ANT2 output pins provide a balanced RF output to the antenna. The manufacturer recommended these pins must have a DC path to VDD_PA, either via a RF choke or via the center point in a dipole antenna. The load impedance seen between the ANT1 and ANT2 outputs should be in the range 200-700Ω. A low load impedance (for instance 50Ω) can be obtained by fitting a simple matching network or a RF transformer. Both loop antenna and 50Ω antenna can be used but loop antenna is chosen for compact, low cost and simple design.

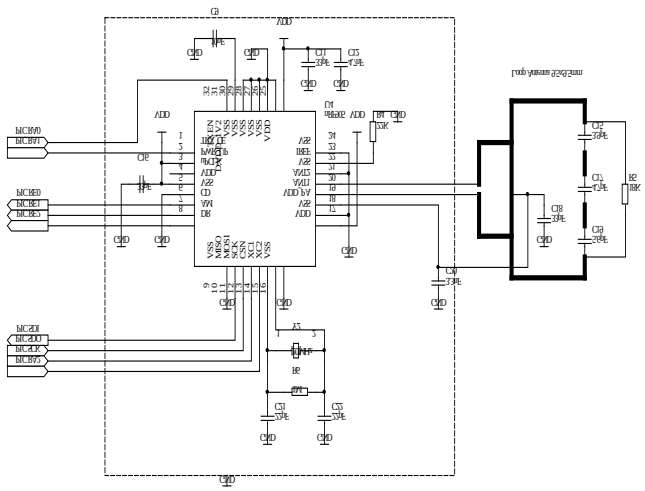


Figure 4. nRF905 Transceiver Section

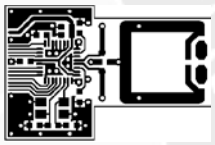


Figure 5. PCB Layout Example for nRF905, Differential Connection to a Loop Antenna (Top View)

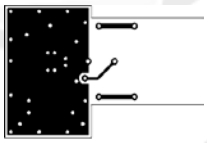


Figure 6. PCB Layout Example for nRF905,Differential Connection to a Loop Antenna (Bottom View)

Figure 5 and Figure 6 show the recommended PCB (Printed Circuit Board) layout for 868MHz operation. The recommended PCB layout is designed for used with SMD (Surface Mounted Device) components and maintains the shortest distance between antenna and components. SMD components cannot available in local and fabrication also needs high end tools. Therefore the recommended PCB layout cannot be used. In this design, only normal components and larger PCB can be used.

Another problem is found for the nRF905. It is a very small IC (5mm X 5mm) and has 32 pins with 8 pins per side. The distance between two adjacent pins is only 0.5mm. The PCB etching is not possible in local facility. Thus, the pins of nRF905 are soldered by small wires by using a special soldering iron and a magnifier and connected these wire to the larger PCB layout. Then the PCB is covered by grounded enclosed tin plate (shown dotted region in the Figure 4) except antenna. The drawback of this poor RF section design is shorter communication range compare with the recommended design.

5.1.2 PIC16LF877A

A 4MHz crystal, Y1, is used for low-voltage operation for PIC16LF877A. Capacitor C_{13} and capacitor C_{14} are crystal oscillator capacitors. Capacitor C_{10} is the decoupling capacitor.

RA0 and RA1 pins control TX_EN and TRX_CE control pins of U4. RB4 pin is configured to send DSR (Data Set Ready) signal to the PC which is connected to T2IN of U1. RB5 pin is configured to receive DTR (Data Terminal Ready) signal from PC from R2OUT of U1.

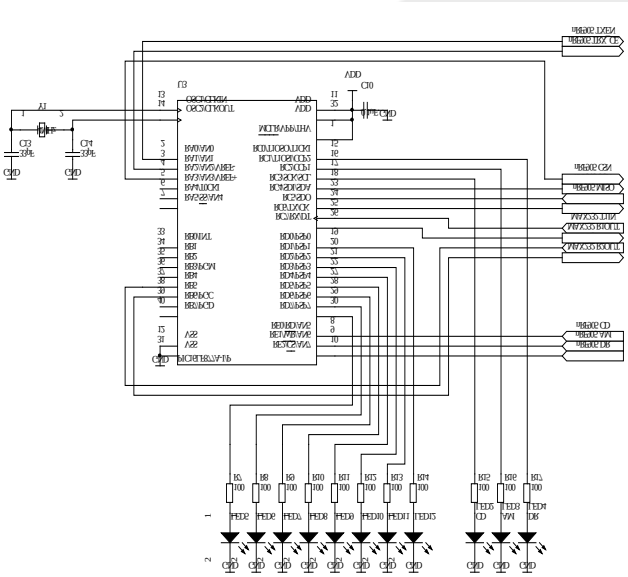


Figure 7. PIC16LF877A section

RC0, RC1 and RC2 pins are connected to 3 LEDs which show DR (Data Ready), AM (Address Match) and CD (Carrier Detect) states read from nRF905 by using port E. SCK, SDI, SDO and RA2 pins are used to interface with nRF905 by using SPI protocol. These pins are connected to MISO, MOSI, SCK and CSN pins of nRF905. TX and RX pins are used to transmit and receive data from RS-232 interfacing with PC by using T1IN and R1IN pins of U1.

All port D pins are connected to 8 LEDs which show transmitted or received characters. RE0, RE1 and RE2 are configured to read CD, AM and DR states from nRF905.

5.1.3 MAX232ACPE

MAX232ACPE RS-232 transceiver from Maxim Integrated Product, U1, is used to interface the microcontroller with the PC. MAX232A has higher slew rate, needs smaller capacitor and faster data transfer rate than normal MAX232. Capacitor C_1 and C_2 are positive charge pump capacitors and capacitor C_3 and C_4 are negative charge pump capacitors. R1IN pin is used to receive RS-232 data from the TX pin (pin 3) of RS-232 port of the PC and inverted these signals to logic level signals which are read by the PIC by using R1OUT pin. T1IN pin reads TX pin of PIC and inverted and send to T1OUT pin which will be connected to TX pin (pin 2) of RS-232 port of the PC. T2IN pin reads from RB4 pin of PIC. It is used to show DSR handshaking signal for the RS-232 protocol. This signal is inverted and appears at T2OUT pin of U1 which is connected to DSR pin of RS-232 port of the PC. R2IN pin reads and inverts the DTR signal from the RS-232 port of the PC and shows at R2OUT pin of U1 which will be read by the RB5 pin of the PIC.

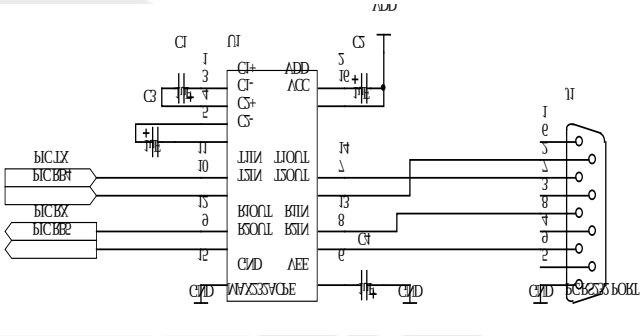


Figure 8. MAX232ACPE Section

Photo of complete circuit and data transferring system are shown in Figure 9 and 10 respectively.

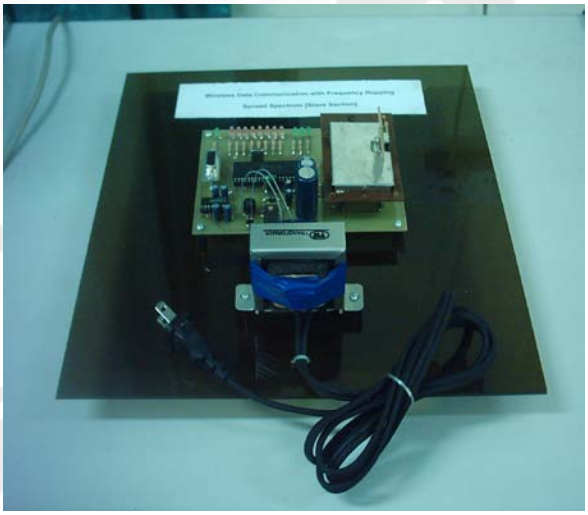


Figure 9. Photo of Complete Circuit

Figure 10. Photo of Data Transferring System

5.2 Software Implementation for the System

The firmware program for the microcontroller is compiled with the PICC Compiler Toolsuite version 8.02 from HI-TECH Software. The source code is written in the MPLAB IDE version 7.01 from Microchip Corporation. The window software for personal computer terminal is written and compiled with Visual C#.NET from Visual Studio 2003.NET from Microsoft Corporation.

The microcontroller uses two communication protocols. It uses UART protocol to interfacing with PC and uses SPI protocol to interfacing with nRF905. In SPI interfacing the microcontroller is in the master mode and nRF905 is in slave mode.

In UART interfacing, the maximum data transfer rate available for the 4MHz PIC, 19.2kbps, is used. In SPI interfacing, the maximum data transfer rate available for the 4MHz PIC, 1Mbps, is also used.

TIMER1 of the microcontroller is used to as the main timer. Each pseudo random frequency channel number is calculated by manipulating of a predefined 16-bit seed number and the 16-bit current clock numbers. The master transceiver is the main clock number generator. The slave transceiver needs to synchronize with the clock number of the master transceiver. The clock number is started from 0 and is increased every 50ms. Each 50ms is called a time slot and the clock number is the time slot number. The master unit transmits its data packet in each even time slot and receives data from the slave unit every odd time slot. The slave unit transmits its data packet in each odd time slot and receives data from the slave unit every even time slot. Every time slot has a number and which is used to generate pseudo random number. Thus, the working frequency is hopping randomly every 50ms. This working frequency must be same on both master and slave unit. The slave unit synchronizes its clock number with the master clock number every time the data packet is received.

In both master and slave transceiver units the odd and even time slots are not identical. To allow for some time slipping, an uncertainty window is defined around the exact receive timing. The window length is 200μs before the exact receive timing. Thus, in master every even time slot length is 49.8ms and every odd time slot length is 50.2ms. In slave every odd time slot length is 49.8ms and every even time slot length is 50.2ms.

The nRF905 transceiver can transmit and receive up to 38 byte data packet each time, including two CRC check bytes, four address bytes and 32 bytes data. In this design, only 24 bytes of 32 bytes data area is used for data. Three bytes from the left 8 bytes is used for two clock bytes and one data count byte.

The most difficult part for the slave transceiver is the time synchronization. The master and slave transceivers must hop their working frequencies at the same time, which means the slave unit needs to adjust its time slot starting point to the time slot starting point of the master transceiver.

The time required for the transmitting of one data packet of nRF905 transceiver is measured by using software timer. It is found that 6.94ms long from TX_CE enable to DR high. If the slave transceiver received DR signal, the slave adjusts its left time for the next time slot to 43.06 (50ms -6.94ms). Thus the starting time for next time slot is synchronized with master. The slave transceiver always adjusts its starting time and time slot number with every received packet.

At first even time slot, the master unit pages for the slave unit. It transmits its time slot number by using a single default frequency. Then it changes working frequency for the receive time slot. At that time, the slave scan for master by stopping its timer, changes the working frequency to the single default frequency and listens to the data packet of the master transceiver. When the data packet is received, the time slot number is set to the master clock number plus one and adjusts its starting time of the next time slot and runs the timer. Then it changes working frequency for the transmit time slot.

In master, at every even time slot start, the master unit transmits its data packet. After transmission, if the transmit packet includes data it send channel number to the PC. Then it changes working frequency and waits for the data from the PC until its time up.

In slave, at every odd time slot start, the master unit transmits its data packet. After transmission, if the transmit packet includes data it send channel number to the PC. Then it change working frequency and wait for the data from the PC until its time up.

At every time slot, the master unit listens to the data packet sent from the slave for 8.2ms. If the data packet is not received before waiting time, it assumes missing packet and increase miss count number. If the miss count greater than 20 the master need to be page to slave and the slave need to be scan for master again. At master, if the data packet is received and the data count byte is greater than one, the master unit sends data to the PC. At slave,if the data packet is received the slave unit updates its time and time slot number.

6. CONCLUSION AND DISCUSSION

The aim of this paper was to obtain a secure data communication transceiver by using FHSS technique. The hardware design includes three major components, a PIC16LF877A microcontroller from Microchip Corporation, an nRF905 RF transceiver from Nordic Semiconductor and MAX232ACPE RS-232 transceiver from Maxim Integrated Product. A personal computer (PC) is also included as the data input output terminal.

The microcontroller is used to send and receive data to and from both PC and nRF905 interface. It uses SPI protocol for interfacing with nRF905 and uses UART SCI protocol for interfacing with PC. MAX232ACPE transceiver works as a voltage level converter between RS-232 level and UART logic level. nRF905 RF transceiver is used as the wireless transceiver module which is configured to run on 20MHz clock and 868MHz working frequency range.

The working frequency channel of this unit is changed randomly every 50ms time slot, which means one random number is generated each time slot. The random number is a 4-bit number which is calculated from microcontroller by using a 16-bit seed number and the 16-bit working time slot number. The seed number is preprogrammed on both master and slave microcontroller. The random number is then added to the base 8-bit frequency channel number to form 16 working frequency channels range from 867.2MHz to 870.2 in 200kHz steps. Both the master and slave unit must generate the same random number at the same time slot to work on same working frequency to maintain the connection. The slave unit needs to check and adjust its clock number with the master clock each time the data packet is received.

The data pack is a packet consisting of 32 bytes. Each data packet consists of two bytes clock number and one byte number of data count in addition to 24 bytes data totally 27 bytes. The other 5 bytes are reserved for further uses. The transceiver transmits one packet in every 100ms and receives one packet in every 100ms. The master unit transmits in each even time slot number receives in each odd time slot number. The slave unit transmits in each odd time slot number and receives in each even slot number. The master always needs to start the connection. It starts by transmitting a paging packet, which consists of the slave address and two clock bytes, by using default frequency. Then the master is receiving from the slave with the random frequency generated from the clock number. The slave unit is always scanning its data by using default frequency every time the master unit is not present. If the address of the transmitted packet matches the slave address, the slave unit reads the clock number in the received packet and synchronizes its clock. Then the slave retransmits the data packet to the master by using the random number generated from the received master clock. Then the connection is established.

The units are tested by using two PCs as data terminal. Testing includes sending characters and text files. The communication range up to 10m can be used and found working well.

6.1 System Benefits

This FHSS transceiver ensures secure data communication. 16 working frequencies changing randomly in every 50ms is difficult to detect. The 32-bit address ensures security and extensibility of the working units. 50kbps data transfer rate is satisfactory for wireless data communication. Three green LEDs indicates the CD (carrier detect) the AM (address match) and DR (data ready) state of the transceiver. The eight red LEDs is used to display the transceiving data. These LEDs visually confirm the connection status. The unit can be used in the area where the noise level is high because it works on multi frequencies. If one frequency is interfered, the next frequency is still working. It can be used especially on military application and the places where the data bandwidth is limited. The slave units can be extended for a large network.

6.2 Further Extensions

The working frequency channels can be extended (up to 2^9 channels) with wideband antenna designs. The well designing of whip antenna can be used for longer communication range. Because of the lack of instruments, the transceiver section cannot be constructed as the recommendation. Therefore the communication range is shorter than mention in the specification. Better data hand-shaking techniques can be designed by using extra packet bytes. The random number generation used in this design is just for testing purpose and it is found that the randomness is not satisfactory. Better random generation technique should be used in the future to get more secure communication. More slave units can be added (up to $(2^{32} -1)$ slave units) to the network by changing communication protocol. Power saving features of both PIC and nRF905 can be used for extended battery life if the unit is designed to operate with 3V lithium battery.

7. REFERENCES

[1] "Basic Communication Theory", INFOSEC Engineering. http://www.blackmagic.com/ses/bruceg/ COMMBOOK/ch1 commbook.html

[2] Dr. Kamilo Feher, "Wireless Digital Communication", Prentice-Hall of India Private Ltd., 2003.

[3] G. R. Cooper and C. D. McGillem, "Modern Communications and Spread Spectrum", New York, Mc-Graw-Hill, 1986.

[4] "Interfacing the Serial/ RS232 Port V5.0" http://www.sent.com.au/~cpeacock

[5] "PIC 16F877A" Datasheet http://www.microchip.com

[6] R. A. Penfold , "An Introduction to PIC Microcontrollers. " London Bermard Babani Ltd., 1997.

[7] R. J. Schoenbeck, "Electronic Communications Modulation and Transmission", Merrill Publishing Company, 1988.

[8] R. L. Peterson, R. E. Ziemer and D. E. Borth, "Introduction to Spread Spectrum Communications", Upper Saddle River, NJ, Prentice Hall, 1995.

[9] Single Chip 433/868/915 MHz Transceiver nRF905. http://www.sparkfun.com/datasheet/IC/nRF905_rev1_0

[10] "Spread spectrum transceiver controlled by PIC", Mikroelectronicka, http://www.mikroelektronika.co.yu/english/Project003/stc1e.htm

[11] S. Rappaport, "Wireless Communications, Principles and Practice", Prentice Hall PTR, 1996.

[12] S. Willian, "Data and Computer Communications", Prentice-Hall, Inc, 1996.

Wireless LAN User Positioning Using Location Fingerprinting and Weighted Distance Inverse

Justinus Andjarwirawan
Petra Christian University
Siwalankerto 121-131
Surabaya 60236
+63-31-2983310
justin@petra.ac.id

Silvia Rostianingsih
Petra Christian University
Siwalankerto 121-131
Surabaya 60236
+62-31-2983455
silvia@petra.ac.id

Charlie Anthony
Petra Christian University
Siwalankerto 121-131
Surabaya 60236
+62-8563097530
ex.animaster@gmail.com

ABSTRACT
Recently, service providers are interested in developing geolocation-based services for users, such as weather dan traffic information using mobile positioning technology. GPS is widely used for determining user's position in an outdoor environment. But GPS lacks reliable performance in an indoor environment.

Because of these reasons, we offer a positioning system application that utilizes wireless LAN which is commonly found in most laptops today. The positioning system uses Location Fingerprinting method which consists of two phases: training and positioning. The objective of training phase is to build a fingerprint database, which is a Received Signal Strength (RSS) data from each installed Access Point. In order to minimize time needed to do site survey, Weighted Distance Inverse interpolation is used. In the positioning phase, user sends RSS data to server and the position is determined based on the fingerprint database.

The result of the experiment shows that the positioning system reliability depends on the fingerprint database. RSS data, which is the main fingerprint data is very sensitive to environment condition. Accuracy of the positioning can be improved by adding more Access Points and determining optimal number of Reference Points. This application can be developed further so that it can cover wider area.

Keywords
Positioning System, Wireless, Networking, Location Fingerprinting, Weighted Distance Inverse.

1. INTRODUCTION
In recent years, mobile positioning is increasingly high demand by telecommunications service providers. This is caused by the increasing number of applications that require accurate location information from mobile devices. In a position outside the building (outdoor), GPS (Global Positioning System) is one technology that can be used to provide users with positioning information accurate enough. But the solution is less appropriate when used in indoor environment. This is caused by weak signals received by a GPS device from the satellites due to impeded by building structures. Along with the increasingly widespread use of wireless LAN system in various indoor environments, such as schools, universities, and shopping centers, arises the idea about using the access point as a solution to provide user position information in the indoor environment [3].

Access point (AP) cannot provide information on the user's position directly [4]. Interconnection is established between the AP and mobile device users, resulting in some information. One of the important information obtained from the interconnection between the AP and the user's mobile device is the Received Signal Strength (RSS) [5]. RSS is said to be important, because this is the only information that can indicate the distance between the AP and mobile device users. Therefore, we need an application that can process information in order to get these RSS users accurate positioning information.

2. OBJECTIVES
The implementation of this problem is based on how to determine the reference point (RP) during field surveys in order to obtain training data that represents every point on the environment being surveyed, and also how to do the interpolation of data using Inverse Distance Weighted method (WDI) to assist in providing reliable training data. The data on AP RSS is stored in a database server in realtime.
The aim of this work is to create a client application to determine the position of connected clients with a wireless LAN in an indoor environment based on information processed by the RSS method and WDI Location fingerprinting.

3. WIRELESS LAN
Basically, wireless LAN (Local Area Network) is a connection between two or more computers without using wires. Wireless LAN technology is the modulation of radio waves to communicate data between devices within a region, called the basic service set. Thus, the client can move the position but remains connected to the network, within the coverage area of the wireless LAN. 802.11 is a standard by IEEE to represent wireless LAN. This protocol has three main layer stacks, they are the Upper Layer, Data Link Layer, and Physical Layer [2]. Authentication like RADIUS is handled above these layers.

Any 802.11a/b/g device can operate at one of the following 4 operating modes:

• Master (AP / Infrastructure)

Wireless LAN in this mode will create a new wireless network name (SSID) and a particular channel. Wireless LAN devices in master mode can only communicate with other devices that run on the mode of managed and associated with it.

• Managed (Client)

Wireless LAN devices in the managed mode will join the wireless network created by a master device and adjust the channel works. In the managed device, generally available RSS information from each access point detected.

• Ad-hoc (Peer-to-Peer)

In ad-hoc mode there are no master or AP devices. Each device connected in ad-hoc network communicate directly with her partner device.

• Monitor

This mode is typically used to monitor the condition of a wireless LAN. In this mode, the wireless LAN device not only sends the data but monitors radio traffic on the specified channel.

4. LOCATION FINGERPRINTING

One method for determining the position of a client using the Wireless LAN is the Location fingerprinting method. This method consists of 2 main stages, namely training phase and positioning. The whole process of location fingerprinting is illustrated in Figure 1:

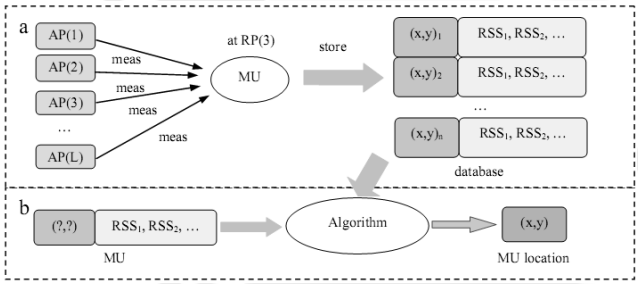


Figure 1. Location Fingerprinting

The purpose of this training phase is to construct a fingerprint database. To produce an optimal database, the selection of Reference Point (RP) needs to be done carefully.

After determining the number of RPs, a client device is placed on an RP and Received Signal Strength measured (RSS) from each Access Point (AP) that was detected by the client. Data obtained from the RP characteristics, namely, the RSS of each AP, and then stored into the database along with the coordinates of the position of the RP. This process is repeated for all of the RPs that have been determined.

In Figure 1, the client described as the Mobile User (MU) and was placed on the RP-3. RSS of AP(1) to AP(2) to the MU are to be measured and the results entered into the database, along with the coordinates of his position, starting from $(x, y)_1$ through $(x, y)_n$.

In the location determination phase, the client at an unknown position will measure the Signal Strength (SS) received from each Access Point [1]. SS measurement results are then compared with the database obtained from the training phase, in order to obtain the client's position.

5. WEIGHTED DISTANCE INVERSE

To save time and effort required in preparing the fingerprint database in training phase, there is a method called data interpolation. Interpolation process is filling the data gaps with data from a particular method of data collection to produce a continuous distribution pattern (Sanjaya, 2006) [7].

One method of data interpolation is a frequently used Weighted Distance Inverse (WDI). This method assumes that each point has a localized effect and its value decreases with distance. The WDI formula is:

$$\hat{Z}(X_0) = \frac{\sum_{i=1}^{n} \frac{1}{d_i^b} Z(X_i)}{\sum_{i=1}^{n} \frac{1}{d_i^b}} \quad (1)$$

Where:
Z = Weight value
b = Value of power factor
X_0 = Location of the point whose value is unknown
X_i = Location of the point whose value is already known
d_i = Distance between X_0 and X_i

6. SYSTEM DESIGN

The development of the system is using Microsoft .NET Framework. Microsoft .NET Framework is a software component that is part of Microsoft Windows. The .NET has a big library and able to meet most needs of programming in various fields, such as user interface, file management (read / write files), database connectivity, network communications, cryptography, web-based application development, numeric algorithms, and so forth.

For the network topology in the system design, in Figure 2 we can see that this mapping system requires several Access Points to be able to determine the position of a client based on RSS [6]. In addition, one of the APs must also be connected to a server via a wired LAN. It is important because it tells the server as its role as a database server, which stores the fingerprint data, client data, and history logs required by the mapping system.

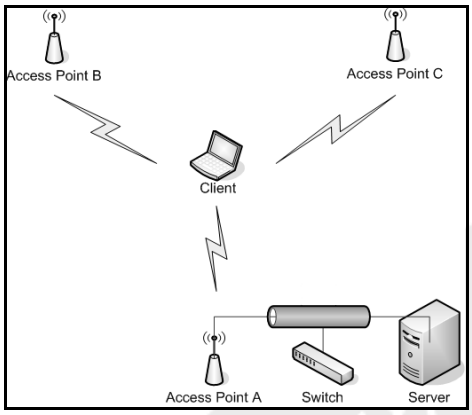


Figure 2. Network Topology

In addition, field surveys are also needed to determine the best positions of APs, so that the mapping system can be implemented on a wider coverage area. Field surveys are also required to collect fingerprint data which is needed in the training stage in the method of mapping by Location Fingerprinting [8].

Generally, system design for this wireless LAN application, a mapping is divided into 3 parts: the design of systems for field surveys, system design for the client, and system design for the server.

6.1 System Design for Site Survey

Application of a field survey was made to assist in collecting fingerprint data, obtained from the RPs which has been determined earlier. In addition, this application can also be used to design the placement of the APs and RPs on the image map used.

Before we can perform a field survey of this system, an image map of the location should be available first. After that, the position of APs and RPs in survey will also be defined. If the survey data already exists, the data can be directly loaded into the application. Conversely, if there is no survey data, the placement of the AP and RP should be defined manually.

The format used to store survey data is the XML data formats. Here is a survey data example:

```
<?xml version="1.0" encoding="utf-8"?>
<datasurvey>
  <accesspoints>
    <AP
      ssid="00:20:A6:84:C2:28 (virtue)"
      nomor="1"
      x="414"
      y="181" />
    <AP
      ssid="00:20:A6:84:BF:FE (dynames)"
      nomor="2"
      x="365"
      y="10" />
  </accesspoints>
  <referencepoints>
    <RP
      nomor="1"
      x="446"
      y="159" />
    <RP
      nomor="2"
      x="380"
      y="159" />
    <RP
      nomor="3"
      x="414"
      y="159" />
  </referencepoints>
</datasurvey>
```

In this XML archive there are two main elements, they are `accesspoints` and `referencepoints`. The `accesspoints` element stores:
`ssid` - the MAC address of Access Point
`nomor` - Access Point sequence number
`x,y` - coordinate position of Access Point

And for the `referencepoints`:
`nomor` - Reference Point sequence number

`x,y` - coordinate position of RP

6.2 System Design for the Client

Client application is used to monitor the APs, records the information of wireless LAN devices that are used by the client, and sends the data to the server to be stored and processed further to determine the client's position.

This system looks for APs that are around the client and the APs' MAC addresses are checked whether among those 3 detected APs have been deployed for mapping systems. When 3 APs have been deployed, then the MAC address and RSS from the AP is recorded and sent to the server periodically.

The system will then match the RSS data that have been sent by the client with the generated fingerprint data. If there is data in accordance with the RSS fingerprint data, the coordinate position (x,y), which correlates with the fingerprint data is inserted into the client as information for the current client's position.

6.3 System Design for the Server

The system displays a map of the location and described the tiny dots on the image to represent the client's current position. Client's position was taken from the client data in the database. The map view is refreshed periodically.

Before performing the data interpolation, the system asked for an input as part of the image map that its RSS data will be generated using the WDI.

After all input values are entered, the system does the RSS data interpolation for points that have not known its RSS, according to the maps.

7. CODING

To begin the implementation, field survey is required to obtain fingerprint data, which forms the RSS data in the training stage in the method of Location Fingerprinting.

For coding the site survey with Microsoft .NET we used a library to gain access to the wireless device, it is called MetaGeek IOCTL NDIS or Input/Output Control Network Driver Interface Specification.

MySQL Connector/Net library is used to access the database system. All information are stored in a MySQL database.

8. EVALUATION

The evaluation takes place in a one floor area of rooms. Two laptops with wireless LAN and 3 access points separated in the area.

Testing is done by holding the notebook in a standing position. Tests conducted on 10 test positions, randomly determined, and at each test position was tested 10 times. In each test, the RSS value is taken at 10 seconds in average.

The next step is to determine the RPs positions and the number of available APs. From these tests, with more number of APs, the determination of the location becomes more stable and accurate. It can be seen from the standard deviation which is lower, the average value of the smaller difference in distance, and higher average value of the validity.

From the results of power factor WDI test, the accuracy of location determination system is good enough when the value of power factor is 3. In the power factor value below 3, the distance between adjacent points is less influence on the prediction signal, resulting in a less than optimal for the data used in location determination system. In contrast to the value of power factor above 3, the distance between adjacent dots increasingly influential, so that the influence of points that is located a little far ignored.

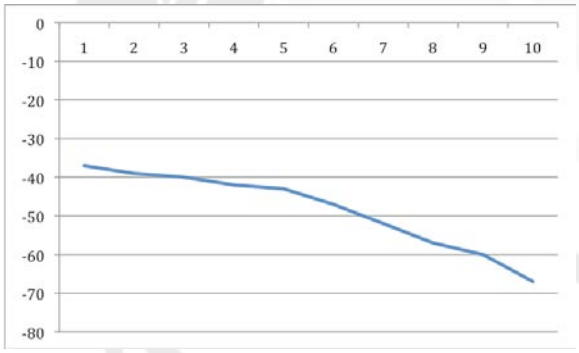


Figure 3. RSS (y axis, in dB) by distance (x axis, in meters)

Figure 3 shows the distance between client and Access Point with its signal strength. From this result, it turns out the relationship between RSS and distance varies, so that relations cannot be determined mathematically.

9. CONCLUSION AND SUGGESTIONS

Based on test results, there are a few things to conclude:

- To produce fingerprint data that represents all points of the position requires the placement of APs and RPs in a field survey to conform the shape of the room and AP's coverage.
- The more APs the more stable and accurate the results are. The added number of APs make the characteristics of each point more unique, because the existing fingerprint pattern also changes.
- To produce an optimum interpolated data, WDI function parameter adjustment is necessary, namely the value of power factor. In addition, the determination of RP was also significantly influence the interpolation, because WDI is using RP as a reference point to interpolate.
- RSS data delivery process from the client to the server could not be done in realtime, because the interconnection between the client and server is via wireless LAN, and that its data transmission speed is strongly influenced by surrounding environmental conditions, such as distance, barriers and level density.
- Depending on the accuracy of location determination of the fingerprint data that is used as a reference, the search algorithm used in fingerprint data, the orientation (direction) notebook and a power antenna on the laptop that are used.
- There is no definite relationship between the RSS and distance. This is because in addition to distance, many other factors that affect the RSS, such as the orientation of the notebook, power antenna, and types of modules used in wireless LANs.
- ActionScript 3.0 and Flash Player display mapping results with the client well enough and responsive.

After an evaluation of the overall system, the authors hope that this work may be developed further with the development of the suggestions as follows:

- Factors affecting variation RSS data received, for example: interception, air temperature, low density and structure of the building, need to be researched more about how big the impact on the quality of fingerprint data generated.
- Parameters determining the location need to be added also to the power antenna and the orientation of the client in order to produce more accurate position.
- To improve the accuracy of location determination, the search algorithm fingerprint data needs to be configured in such a way that it can recognize patterns in the fingerprint database.
- The system can also be developed so that not only it can detect the client's position that is on the same floor, but also the different floors and a wider range of location determination.
- This location determination systems can also be extended with location-based application service provider, such as weather information, event information, etc.

10. REFERENCES

[1] Bahl, P., & Padmanabhan, V. 2000. RADAR: An in-building RF-based user location and tracking system. Proceeding of 19th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM-2000), 2, 775–784.

[2] IEEE Computer Society. 2007. IEEE Standard for Information technology Telecommunications and information exchange between systems - Local and metropolitan area networks specific requirements - Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications. 25 April 2008. http://standards.ieee.org/getieee802/802.11.html

[3] Jan, R. & Lee, R. 2003. An indoor geolocation system for wireless LANs. Proceedings of the 2003 International Conference on Parallel Processing Workshops (ICPPW'03), 1530-2016/03

[4] Kontkanen, P., Myllymaki, P., Roos, T., Tirri, H., Valtonen, K., & Wettig, H. 2005. Topics in probabilistic location estimation in wireless networks. 5 Maret 2008. http://cosco.hiit.fi/Articles/

[5] Li, B., Salter, J., Dempster, A., & Rizos, C. (2007). Indoor positioning techniques based on wireless LAN. 10 Februari 2008. http://www.gmat.unsw.edu.au/snap/publications/

[6] Roos, T., Myllymaki, P., Tirri, H., Misikangas, P., & Sievanan, J. 2002. A probabilistic approach to WLAN user location estimation. International Journal of Wireless Information Networks, 9(3).

[7] Sanjaya, Hartanto. 2006. Spatial analyst: interpolasi grid dari data titik. 10 Maret, 2008. http://hartanto.wordpress.com/2006/04/06/sa-interpolasi-grid-dari-data-titik

[8] Shepard, Donald. 1968. A two-dimensional interpolation function for irregularly-spaced data. Proceedings of the 1968 ACM National Conference, 517–524

Wlanxchange : A New Approach in Data Transfer for Mobile Phone Environment

Ary Mazharuddin Shiddiqi
Informatics Department, Faculty of Information Technology, Sepuluh Nopember Institute of Technology
ary.shiddiqi@cs.its.ac.id

Bagus Jati Santoso
Informatics Department, Faculty of Information Technology, Sepuluh Nopember Institute of Technology
bagus@cs.its.ac.id

Rio Indra Maulana
Informatics Department, Faculty of Information Technology, Sepuluh Nopember Institute of Technology
rim1910@gmail.com

ABSTRACT

The use of Wireless LAN technology has in mobile phone products has attracted more research in this technology. One of the most common uses of WLAN is to connect to the internet via local area networks.

As the advance of mobile phone technology march out, the WLAN technology is now available on recent mobile phones products. Preliminary, the mobile phone WLAN is used commonly to connect to local area networks (LAN) as its legacy use. However, there is an opportunity that this technology can be used for file transfers between mobile phones and other network-based devices.

This research proposes an extended use of the mobile phone WLAN in the data transmission between mobile phones and other network-based devices. Experiments results show that the data transmission can be used to transfer files among network-based devices with high level of satisfactory.

Keywords

PyS60, WLAN, Data Transfer.

1. INTRODUCTION

The most commonly used operating system in mobile phones is Symbian 60. On this platform, data transfer application commonly uses Bluetooth technology due to its cheap production cost and it has been widely installed on most mobile phone types.

In the previous research [5], Shiddiqi et al. has developed an application that can transfer files between mobile phones using Wireless LAN technology. This research extends the previous research by conducting more research on this application to explore the used of the mobile phone WLAN technology and names the application as WLANXChange.

2. PAGE SIZE

Wireless LAN (WLAN) has components that build the architecture of the WLAN network. The components consist of station, basic service station, and extended service set. The details of the components are as follows:

2.1 Station

Station is a component in WLAN architecture known as node. All components that can be connected through a wireless medium in a WLAN are called a station. All stations are equipped with wireless network interface cards (WNICs). Wireless clients can be mobile devices such as laptops, personal digital assistants, IP phones, or PCs that have been equipped with wireless network interface.3

2.2 Basic Set Service

Basic service set (BSS) is a collection of stations that can communicate between each other. There are two types of BSS, i.e. the Independent BSS and Infrastructure BSS. Each BSS has a BSSID obtained from the MAC address of the access point serving BSS. An independent BSS is an Ad-Hoc network that does not have an access point, so that the BSS is not able to connect with other BSS. As for infrastructure BSS may be associated with other stations that are located on the same single BSS through access point.

2.3 Extended Set Service

Extended service set (ESS) is a set of connected BSS. Access points connected to the ESS distribution system. Each ESS has an ID called the SSID is a 32-byte (maximum) character string.

3. PYTHON FOR SYMBIAN 60 (PYS60)

Python for Symbian S60 (Pys60) is the development of specific language used python for the Symbian smart phone platform Series 60. In addition to the standard features of the python language, PyS60 provides access to many unique functions of smart phones such as camera, contacts, calendar, audio recorder, games, TCP / IP and Bluetooth communication simple.

PyS60 is open source, under Apache 2 license and python. Python for S60 based on the Python version 2.2.2. not only supports many of the modules Python Standard Library, but also includes some special modules for mobile platforms, such as Native GUI elements, Bluetooth, networking, GSM Location Information, SMS messaging, access to the camera, and others. Nokia makes python bindings for Symbian OS is the public API provided on the device S60.

4. SYSTEM ARCHITECTURE

To enable the WLANXChange to transfer files, the first thing to do is to connect to the WLAN network either in Ad-Hoc and infrastructure. In addition, to facilitate the user in recognizing connected devices in a network, we use the device identification system ID. To provide flexibility in user interaction, it would require a communication between devices where each device can communicate with each other and both were doing the sending request, the denial request, receiving the request, and the user can send data to other devices that run the same applications on a

network. Figure 1 shows the components needed to make the technique works.

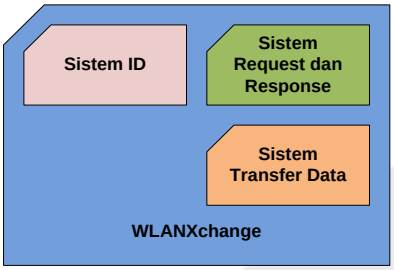


Figure 1. WLANXchange components

4.1 System ID

ID used to identify other devices that run the same applications in a single subnet. ID of each device will be sent via a packet data transport protocols using UDP (User Datagram Protocol).Transport protocol UDP is selected as this ID packet data is not large. This protocol is suitable for use in the recognition process ID from a cell phone to be connected because the search process ID must be updated should there is a change in the device to each node belonging to the subnet. Figure 2 shows the packet format string ID.

Figure 2. Format string packet ID

STATUS can be filled by binary codes 0 or 1, 1 indicates that the device is in a state ready to receive requests or is not carrying out a process of both sending and receiving. Meanwhile, if the status value 0 indicates that the device is in a state could be due to busy doing the sending or receiving data. Figure 3 shows an example of packet format string ID.

Figure 3. Example of format *string* packet ID

Those examples explain that the status of the device named 5800 XM is ready to receive and send data.

4.2 ID Receiver

ID receiver in charge of receiving data packets from the device ID other in a network, and enter into the list of devices and update them. List of device is the collection ID of the device using the same application. Application Receiver ID will bind and listen on port 7654 address

4.3 System Request and Response

Request is a set of commands used two devices to perform handshaking before data is sent. More precisely this request is a handshaking protocol. Figure 4 is a list of the request and the response code of this protocol.

code	format message
1	total_files[space]total_size[space] sender_addr
2	sender_addr
3	null
4	null

Code	Explanation
1	This code is sent when a device wants to send to another device. For example:  It means that a device (192.168.0.5) wants to send 5 files with total size of 35000000 bytes.
2	This code is a response of code 1. This code is sent from receiving device to sending device. For example :  It means that a device (192.168.0.7) is ready to receive files from sending device.
3	This code is used to reject file transfer request from sender.
4	Code 4 is used when receiving device's storage is not enough to store file(s) that will be transmitted.

Figure 4. List of the request and the response code from the handshaking protocol

To be able to do the sending and receiving the request and response, it can take 2 pieces of software components in the receiver and sender Request.

a. Receiver

Just like an ordinary receiver, receiver served as the entrance to receive and process packets in accordance with the request in the specified format and then display it to users

b. Sender

Sender will send packets to the sender requests that the on going. In addition, to send request packet, the sender also sends a good response-response is determined by the user and will automatically send the software.

Both of request and response packets are sent using the transport control protocol (TCP). This is because these packages are very important and must be guaranteed to be sent to the recipient to the back, and make the next routine. Request and response running on port 6543

4.4 Data Transfer System

The data stream is sent in the data stream that runs on the TCP transport protocol. Stream packets of data arranged as in Figure 5.

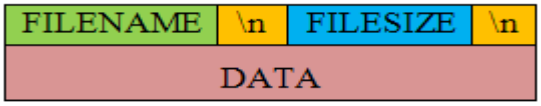


Figure 5. The format of data streams packets

Name of the file is sent with the intention of naming files on the recipient, but if there is the same name in the destination folder on the recipient, the name of the file is automatically generated by the software. Packet data transmission channel will only open when the handshake of the two devices has been formed (code 1 in the back with a code 2). When the recipient code provides 2 replies to the sender, the recipient will open the 5432 port, and the device is ready to receive data packets. Once the code is received, the device 2 will start to send the data. Then the device will open the 5432 port channel to a receiver to send data and the channel will immediately ends when all data has been completed is sent once the port is closed.

5. EXPERIMENTS

This section will discuss both functionality testing and performance of WLANXchange in data transfer between phones using WLAN transmission media. The scope of this test are as follows:

- Two phones that run the Symbian operating system 60 5th Edition and 1 pc phone operating system Symbian 60 3rd Edition
- Python 1.9.7 for each phone and recompiled python shell scripts that have been signed for each phone
- Using 3 phone specifications respectively as follows:
  - 2 Nokia 5800 XM, with 128 MB of RAM and 7 GB of storage
  - 1 Nokia E51, with 96 MB RAM and 256 MB of storage

5.1 Performance Testing

This test is intended to find application performance in data transfer. Performance is measured in this test is the transfer rate and maximum distance.

5.1.1 Transfer Rate Testing

There are several variables of transfer rate on this test that are changed to determine the effect of the variable transfer rate.

- **Bluetooth vs. WLAN**

This scenario compares the speed of 2 different connection modes of Bluetooth and WLAN.

- **Single File vs. Multiple File.**

In this scenario compared between sending files one-one with sending multiple files. However, multiple delivery modes used files and queue files are not sent at the same time.

Transfer rate tests conducted at all the variables are at best performance. That is the distance used is 1 meter. Test results shown in Table 1.

Table 1. Testing speed data transmission between WLAN vs Bluetooth

No	WLAN		Bluetooth	
	Single	Multiple	Single	Multiple
1	423.90	416.00	83.00	84.90
2	382.40	393.60	83.10	82.90
3	371.10	371.80	82.70	83.00
4	388.20	392.70	82.10	84.00
5	368.50	399.20	83.10	85.20
AVG	386.82	394.66	82.80	84.00

From the results of these tests, (units in Kbps) shows that the use of the WLAN connection has a transfer rate which is greater than Bluetooth. The speed of data transmission for WLAN reaches 5 times the speed of Bluetooth at the same distance.

5.1.2 Maximum Distance Testing

In this scenario, the maximum distance calculation experiments carried out by comparing between a WLAN connections with a Bluetooth connection. Each distance is measured by the distance that lies between the two devices in all tests done on an open space with no obstacles between the devices with other devices. The results of testing are shown in Table 2 and Table 3.

Table 2. Testing the maximum distance in WLAN

Jarak(m)	Datarate(Kbps)					AVG (Kbps)
	1	2	3	4	5	
2	505.6	502.0	472.8	468.7	463.8	482.6
4	503.9	499.4	384.6	466.1	465.4	463.9
6	466.8	462.7	469.2	473.0	328.0	439.9
8	499.3	498.4	490.2	478.8	494.5	492.2
10	473.1	424.3	482.8	472.1	453.0	461.1
12	478.7	503.6	520.5	475.1	479.0	491.4
14	451.6	477.7	495.8	345.6	502.9	454.7
16	406.4	513.0	515.7	478.6	483.7	479.5
18	503.9	303.7	502.6	475.4	474.2	452.0
20	464.2	462.2	460.0	492.0	501.7	476.0
22	427.0	442.0	427.3	428.4	508.9	446.7

Table 3. Testing the maximum distance on the Bluetooth

Jarak (m)	Data Rate (Kbps)					AVG (Kbps)
	1	2	3	4	5	
2	83.00	82.10	83.00	83.10	83.00	82.84
4	75.00	77.00	74.20	76.10	71.60	74.78
6	0.00	0.00	0.00	0.00	0.00	0.00

From the table it shows that the maximum distance a WLAN connection farther than Bluetooth, which is 22 feet compared with a Bluetooth connection which can only transmit data at a maximum distance of 4 meters.

5.2 Indirect File Transfers

This scenario is aimed to test the WLANXChange reliability in transferring files using an access point as a bridge of file transfers. The access point can also be used to extend maximum distance that can be covered between transferring devices. The mechanism of connecting mobile phones through an access point can be shown in figure 6.

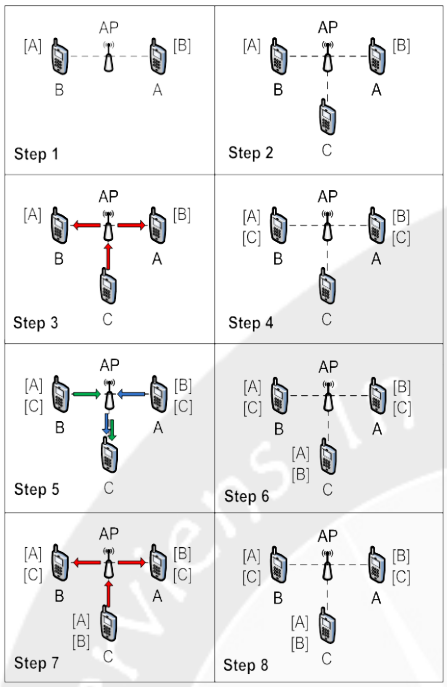


Figure 6. The Mechanism of indirect connection

As shown in Figure 6, the process of file transfer stars by connection the sending and receiving device to the nearest access point (step 1). If a mobile phone wants to transfer a file to one or two devices, it broadcasts cell id request to all connected mobile phones connected to the access point (step 3). Upon receiving the cell id request, the receiving mobile phones reply the cell id request by sending back response packet to the requesting cell id mobile phone (step 5). Having the destination cell id(s), the sending mobile phone sends the file(s) to the destination mobile phones (step 7).

This test used an access point with IEEE 802.11g standard that has 100 MBps throughput. The reason of using this type of standard is to avoid bottle neck problem caused by lower throughput capacity of access point being used. The result of the experiment is shown in table 4.

Table 4. Transfer Rate of Indirect Connection

Distance(m)	Datarate(Kbps)					AVG (Kbps)
	1	2	3	4	5	
2	469.2	498.7	462.4	466.8	460.5	**471.5**
4	494.5	485.2	462.7	436.1	478.7	**471.4**
6	460.5	387.9	479.2	481.1	372.9	**436.3**
8	469.2	476.1	480.2	442.0	432.8	**460.1**
10	427.9	382.7	424.3	503.9	453.0	**438.4**
12	392.1	390.8	385.9	530.1	468.2	**433.4**
14	423.5	390.5	487.1	345.6	382.1	**405.8**
16	339.2	450.1	323.2	444.1	401.5	**391.6**
18	322.2	286.7	345.1	299.1	391.2	**328.9**
20	314.2	299.8	321.4	399.2	384.6	**343.8**
22	303.3	297.2	321.3	351.1	333.1	**321.2**
24	150.7	238.7	194.5	284.2	222.2	**218.1**
26	175.6	265.3	199.8	254.8	299.3	**239.0**
28	201.4	188.8	250.4	211.9	195.5	**209.6**
30	166.6	175.7	143.1	183.0	144.9	**162.7**
32	143.7	149.5	160.9	131.2	137.7	**144.6**

The table shows that there is a decrease in performance of transferring throughput compared to the direct transferring in the same distances. This is due to the use of WLAN access point takes a few seconds to process file request and response from source and destination between two sending/receiving devices. By comparing direct vs indirect transferring experiments, it can be observed the relations of distance and transferring throughput. By increasing the distance of two meters, the average transferring throughput decreased 4.3 Mbytes or ± 1% of direct connection at the same distance.

However, there is a positive side of using access point. As shown in table 4, the maximum distance that can be reached of the sending/receiving devices incdreased. As shown in table 4, the maximum distance of indirect connection is 32 meters. This is 10 meters further than of the direct connection which is only 22 meters (as shown in table 3). Thus, the throughput decrease caused by the use of access point is patched-up by the increas of maximum transferring distance.

5.3 Transferring to Computer

This scenario observed the transferring performance of mobile device to computer by using the same WLANXChange framework. In this scenario, a mobile device (Nokia 5800 XM) and a laptop were used. The laptop used in this scenario has IEEE 802.11g of its WLAN standard. This experiment was conducted in the same distance measurements. The result of this experiment can be seen in Table 5.

Table 5. Transfer Rate from 5800 XM to Laptop

Distance(m)	Datarate(Kbps)					AVG (Kbps)
	1	2	3	4	5	
2	551.6	513.5	520.9	550.0	533.5	**533.9**
4	530.3	544.6	511.1	512.5	507.6	**521.2**
6	457.4	444.7	434.3	420.1	426.2	**436.5**
8	429.5	423.9	421.3	420.0	416.6	**422.3**
10	348.2	350.8	312.7	333.3	350.1	**339.0**
12	347.9	339.5	340.8	328.9	333.3	**338.1**

The table shows that at the initial distance (2 meters), the transfer rate is higher than direct connection between two mobile devices. This is shown from 2 meters up to 4 meters. However, increasing the distance from this point forward will cause the decrease performance compared to direct connection between two mobile phones. On the average, the decrease is 60 Mbps per 2 meters additional distance.

5.4 Bottle Neck Problem

Once again, this research conducted an experiment regarding bottle neck issue. The experiment used the same devices to test the bottle neck problem. The test was done by using Nokia 5800 XM with the Nokia E51. This is because Nokia E51 (20 MB RAM) has lower computing capacity than Nokia 5800 XM. In addition, when receiving a stream file, the receiving device has to receive file and write the received file to local storage. This is the reason why the receiving device needs large resources to run the two tasks simultaneously.

In this experiment, files with larger size were used. This approach is aimed to test if there is an effect when using small size files compared to one large size file. The result of the experiment can be seen in table 4 and 5.

Table 4. Transfer Rate from 5800 XM to E51

No	Datarate (Kbps)
1	15.8
2	14.3
3	13.3
4	12.5
5	13.2
AVG	13.8

Table 5. Transfer Rate from E51 to 5800 XM

No	Datarate (Kbps)
1	321.4
2	315.4
3	341.5
4	220.3
5	214.2
AVG	282.56

The table 4 and 5 shows that transferring a larger file size produced higher data rate compared to transferring a number of files with similar total size [5]. This can be caused by the process of sending a number of files need extra steps to complete, i.e. selecting file and start sending after the previous file successfully sent.

6. CONCLUSION

The proposed technique of sending file using WLAN on mobile phones is valid and feasible to implement. The experiments show that by using WLAN, the file transfer between mobile phones is more reliable than using bluetoth. This is proven by its speed and connection reliability which were faster and produced higher throughput than bluetooth. Another thing to consider is that by using the WLANXChange, the users do not have to open their inbox if they want to receive files as in bluetooth technology. The users can directly store their received files to their storage on their mobile phones.

7. FUTURE WORKS

It may be a good for this application to have its own protocol, because the current application uses TCP/IP. The dedicated protocol will increase the transfer rate of the application.

Another way to increase the ability of this application is by changing the transfer technique by making this application able to transfer files simultaneously. This can be done if the mobile phone has large memory size.

8. REFERENCES

[1] Railfans, Hedwigus. 2008. S60 5th Edition. (Online), (http://www.hedwigus.com/s60-5th-edition/ , diakses 25 Desember 2009)..

[2] Prasimax Research Division. 2002. Protokol TCP/IP Bag. 1. Depok: Prasimax.

[3] Anonim. 2005. Wireless LAN. (Online), (http://en.wikipedia.org/wiki/Wireless_LAN , diakses 25 Desember 2009)

[4] Fachruzi, Ardhi. 2009. Design Build Client-Server Applications in Mobile Network-based Symbian S60 (PyS60) With Webcam Server, ITS SURABAYA.

[5] Shiddiqi, AM, Santoso, BJ, Maulana, RI, Pembuatan Aplikasi Transfer File Antar Perangkat Mobile Melalui Media Jaringan Nirkabel Berbasis Pys60, Seminar Nasional Teknologi Informasi Dan Aplikasinya (Sentia'10), Politeknik Negeri Malang

Analysis Influence Internal Factors on Fuzzy Type 2 Performance of Swing Phase Gait Restoration

Hendi Wicaksono
Electrical Engineering Dept. Universitas Surabaya
Raya Kalirungkut, Surabaya, 60293, Indonesia
+62312981157, ext: 86.
hendi@ubaya.ac.id

ABSTRACT

We knows from [1] that Fuzzy Type 2 controller can solve stability criterion problem better than Ordinary Fuzzy or we called on this paper with Fuzzy Type 1. On [1] shows on experiment results graphics that Fuzzy Type 2 more stable than Fuzzy Type 1. From [2], the experiment results show that Fuzzy controller can be a good controller than PID controller, but it still shows there is an oscillation effect. In [2] conclusion that Fuzzy controller to be an effective to implement the cycle-to-cycle method on human gait control. We do some experiment with several normal subjects to analyze performance Fuzzy Type 2 control implement on Functional Electrical Stimulation for swing phase gait restoration. In this paper, we describe that Fuzzy Type 2 can control swing phase gait restoration with minimum oscillation than Fuzzy Type 1 on [2] experiment. We also describe that there are an internal factors which can disturb control action. They are a muscle fatigue and muscle force potential. We can analyze which one on experiments that an internal factor appears and make Fuzzy Type 2 must be changed their value.

Keywords

Fuzzy Type 2, Swing Phase Gait Restoration, FES, Cycle-to-Cycle, Muscle Fatigue, Muscle Force Potential.

1. INTRODUCTION

Human gait have 2 phases, stand phase and swing phase [2]. Swing phase gait need a muscle ability to make sure perfectly gait process. Even we got SCI (Spinal Cord Injury), we can lost a muscle ability. For that purpose, we research about swing gait phase restoration. My research continue the previously research [2] about swing phase gait restoration. I was develop a system control for minimize an oscillation level which still happened on result of the previously research which used Fuzzy Controller. Karnik and Mendel introduced Fuzzy Logic Type 2 [3]. A previously Fuzzy Logic called with Ordinary Fuzzy or Fuzzy Type 1. Many systems which have a non stationer noise measurement have an uncertainty linguistic. This make an imprecision level exceed a fuzziness level of Ordinary Fuzzy. Fuzzy Type 2 have a vagueness and uncertainty level higher than Ordinary Fuzzy, because Fuzzy Type 2 have a membership grade as a linguistic form not the same with Ordinary Fuzzy which have a membership grade as a real value. Beside that, Fuzzy Type 2 have a Footprint of Uncertainty (Figure 1) make this control have an imprecision and uncertainty level higher than Ordinary Fuzzy or Adaptive Fuzzy. Usually, Fuzzy Type 2 used on solve the noise a measurements from a sensors [4]. Until the day, there is no research have been found a sensors with zero noise to measure angles of human gait.

In this research, we build a FES stimulator, open-loop test FES stimulator with several subjects, design of Fuzzy Type 2 Controller, closed-loop test with Fuzzy Type 2 controller, analyze a control performance of Fuzzy Type 2 controller. A block diagram system as seen as a Figure 2. Block diagram of Fuzzy Type 2 as seen as Figure 3, and block diagram of plant as seen as Figure 4.

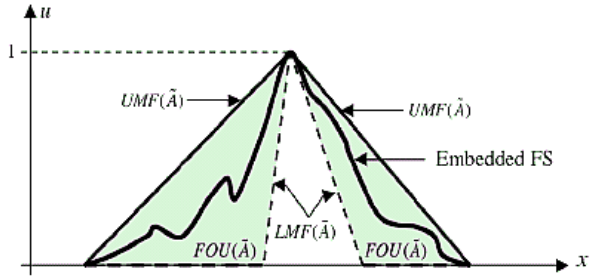


Figure 1. Footprint of uncertainty (FOU) [4].

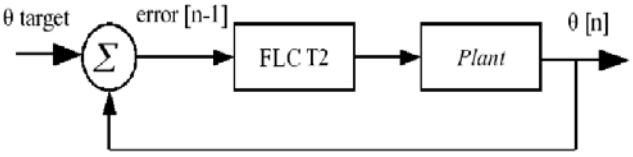


Figure 2. Block diagram [4].

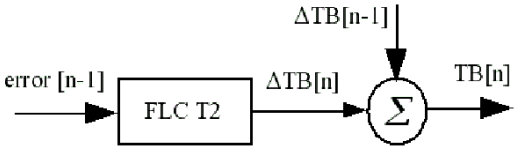


Figure 3. Block diagram fuzzy type 2 [4].

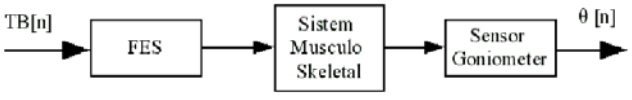


Figure 4. Block diagram of plant [4].

We can see from block diagram that input Fuzzy Type 2 controller is an error [n-1], the output Fuzzy Type 2 controller is a delta Time Burst (ΔTB). This TB[n] which control of FES duration to control how long stimulate will be active to stimulate musculo-skeletal systems. FES must be connected to some electrode and impressed on skin with specific muscle under skin. It's can see on Figure 5.

Fuzzy Type 2 on General Form can not be realized in a real time, because it is very complex calculation so it takes more times. For this problem, a Fast Geometric Defuzzification Method [9] was used on this research. It is can use on a real time, but a level of vagueness and uncertainty is limited to a General Form Fuzzy Type 2. With Geometric model, Fuzzy Type 2 divided into 5 areas a big triangle and the final form will be seen as a polyhedron geometry. It can see on Figure 6. Fuzzy Type 2 was design on a computer for easy calculation using high language. On Figure 7 we can see a computer simulation test of Fuzzy Type 2 Controller. More details of design and experiment result Fuzzy Type 2 was published on SNTF Regional Conference [5].

Figure 5. Electrode position for vastus and hamstring muscle.

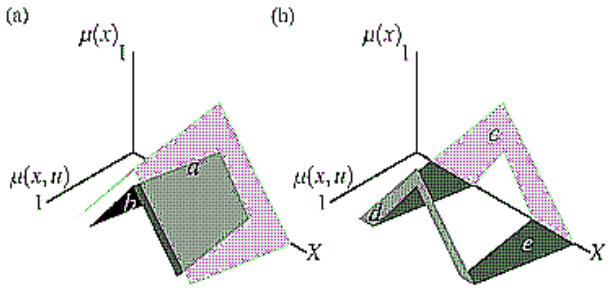


Figure 6. Fuzzy type 2 divided into 5 areas [4].

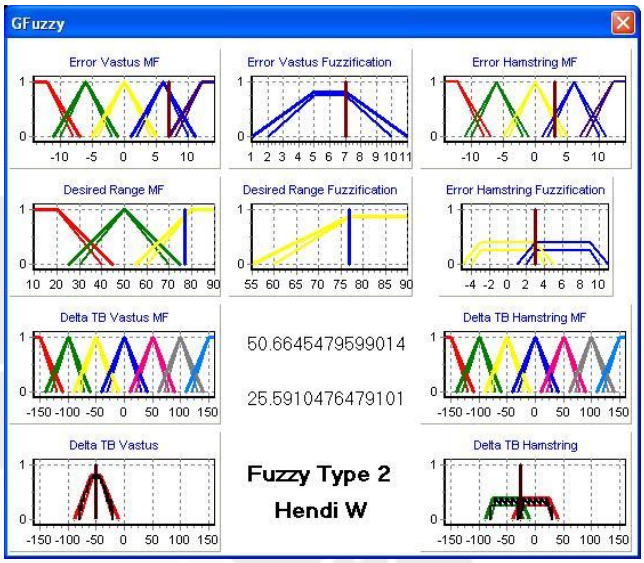


Figure 7. Simulation test of fuzzy type 2 [4].

2. CORRELATION VALUE

Internal factors influence a system control can be check with correlation and dependencies value. We knows that correlation as show as Equation 1.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}, \quad (1)$$

In statistics concept, correlation are statistical relationships between two or more random observed data values. For an example are correlation between demand for one product and its price. Correlations are useful because they can predict a relationship that can be exploited in practice.

Correlation value between absolute value 0 to 0.6 indicate that output control does not influenced by an external factors control. Correlation value between 0.6 to 1 indicate that output control influenced by an external factors control. An external factors control in this research defined by two internal factors as a muscle fatigue and muscle force potential. Muscle fatigue defined that muscle can be under contraction compare with previously angle, its mean that the angle produced lower than previously angle. Muscle force potential defined that muscle can be over contraction compare with previously angle.

3. EXPERIMENT RESULTS

Experiment was taken by a normal subjects. In one experiment, each subject got 30 cycles and the data recorded automatically with computer. One experiment taken approximately 5 until 10 minutes. After got 30 cycles, subject must be have a rest time about 30 minutes. After got a rest time, we can do an experiment with the same subject. We do five times experiment for each subject, but only two experiments show on this paper.

There are two actions, first action is knee extension which vastus muscle was stimulated, second action is knee flexion which hamstring muscle was stimulated. Target knee extension on 40

degree, an target knee flexion on 90 degree. Each chart must be show two lines, upper line is knee flexion representation, and bottom line is knee extension, except correlation charts. On each knee join angle chart, it is also show a dash line which indicate oscillation tolerate about ± 5 degree. It means that if the knee extension angle still on 35 until 45 degree, its still on the track, no oscillation. For knee flexion must be fit on 85 until 95 degree, for called no oscillation.

For each figures, show six charts, first chart is a knee joint angle (deg) chart, second chart is a Time Burst duration (sec) chart, third chart is an error knee joint (deg) chart, fourth chart is a sensitivity (deg/sec) chart, fifth chart is a correlation chart for vastus muscle chart, and the last is a correlation chart for hamstring muscle chart.

In this paper does not show the calculation tables because it is need a more spaces, but only show a chart, it is enough to be see a correlation values representation.

3.1 Subject 1.

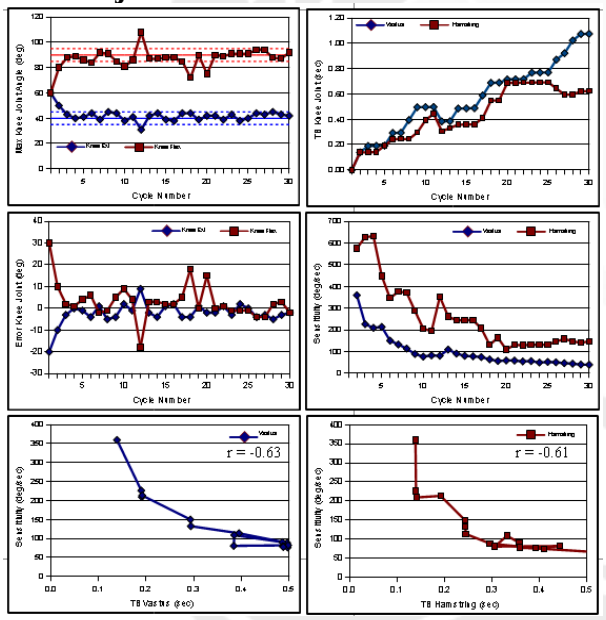


Figure 8. Experiment 1 on subject 1.

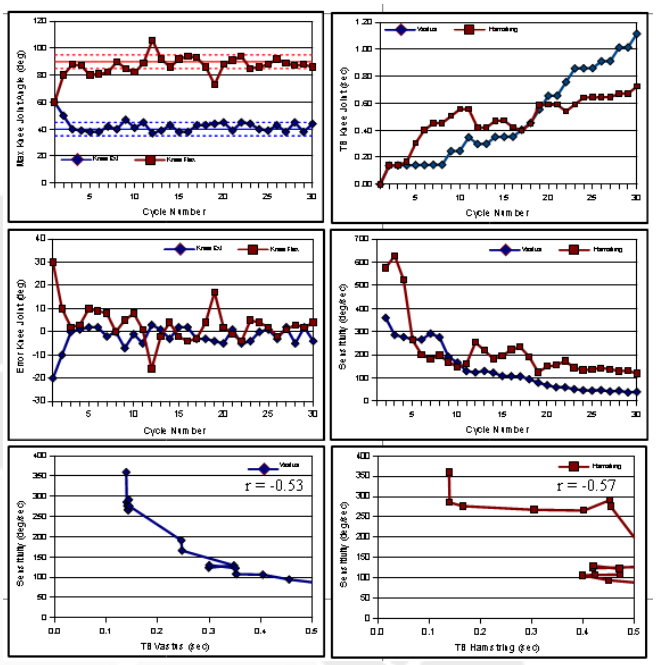


Figure 9. Experiment 2 on subject 1.

3.2 Subject 2.

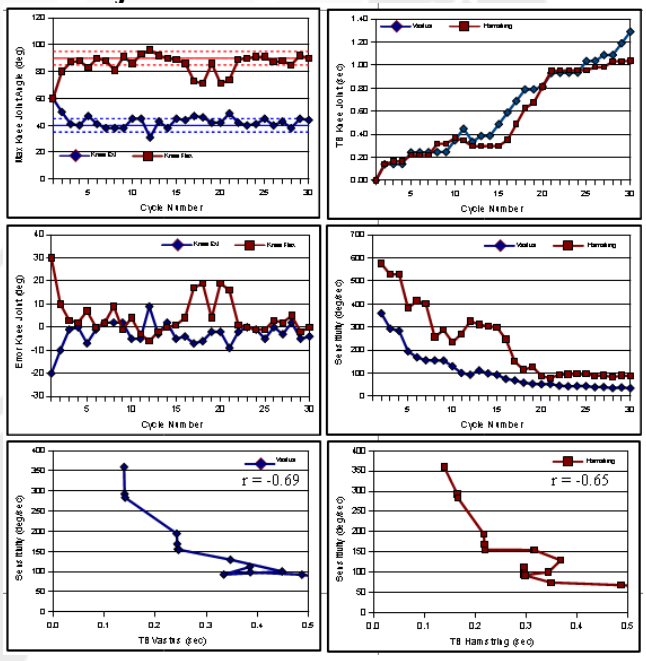


Figure 10. Experiment 1 on Subject 2.

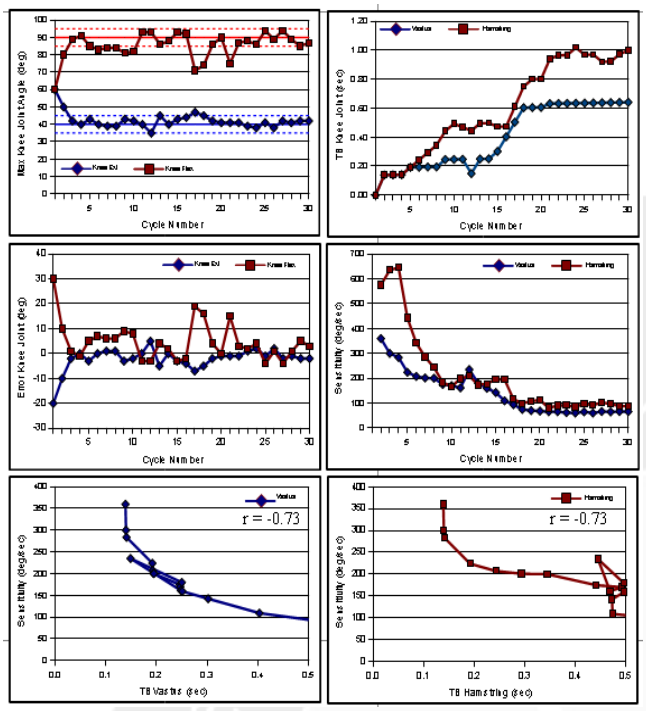


Figure 11. Experiment 2 on subject 2.

3.3 Subject 3.

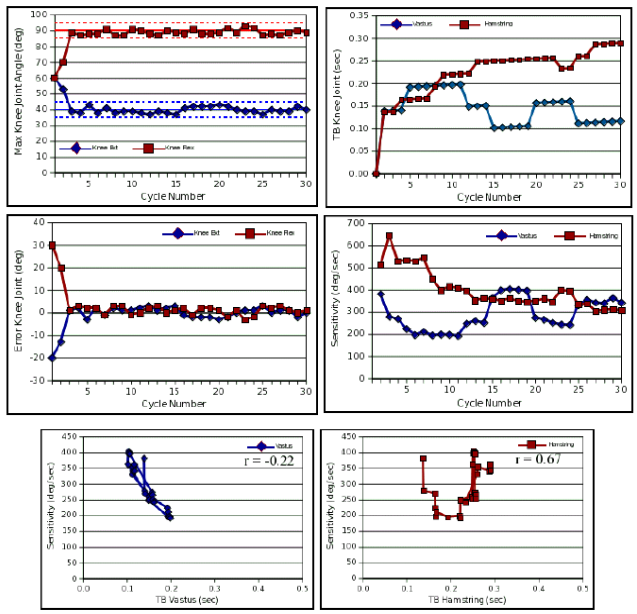


Figure 12. Experiment 1 on subject 3.

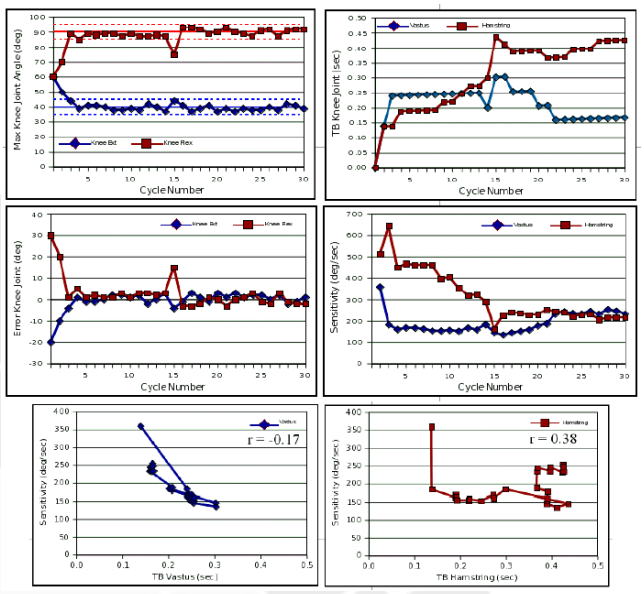


Figure 13. Experiment 2 on subject 3.

3.4 Discussions

Sensitivity defined with difference between angle joint in degree with Time Burst duration (TB) in second. Correlation value between sensitivity (deg/sec) with TB duration (sec). In Figure 8, we can see first experiment result on subject 1. Correlation value (r) is more than absolute 0.6. It's indicate that the oscillation caused by external parameter of control. On cycle 12^{th} for hamstring muscle, muscle force potential found there, and cycle 18^{th} and 20^{th} for hamstring muscle, muscle fatigue found there. On Figure 9, we see the r value quite less than 0.6, but it is limited to 0.6, so, we can conclude that there are an external parameter control but it's not dominant.

For subject 2, if we see from a correlation value that indicate all experiment influenced by an external parameter. Subject 2 have a muscle fatigue more than 1 cycle, on 18^{th} and 21^{th} in experiment 1, and 17^{th} and 20^{th} in experiment 2, but there is not a muscle force potential.

On Figure 12 and Figure 13, we can see two experiments of subject 3 that the Fuzzy Type 2 working perfectly with minimum oscillation ± 5 degree. The value of correlation very low, limited to 0, it's indicate this experiment does not influenced by an external parameter.

All of experiments which analysis found muscle fatigue and muscle force potential, we can see a responses of Fuzzy Type 2 is very fast. It's not need more than 2 cycles to change their values and recover a muscles contraction.

4. CONCLUSIONS

Correlation value can be a method to detection external control factors. With statistic equation, we can measure what the value between sensitivity and TB duration. Actually we got from experiment that the oscillation happen caused by muscle fatigue and muscle force potential have a correlation value between 0.6

until 0.8. Also we got from experiment, a correlation value which muscle fatigue and muscle force potential does not happen between 0.15 until 0.57 in absolute value.

Minimal oscillation from Fuzzy Type 2 about under ± 5 degree. Minimal oscillation happen with there is no correlation between sensitivity and TB duration. On [4], there is a conclusion that the oscillation level Fuzzy Type 2 controller got minimum than Ordinary Fuzzy controller.

From experiments we know that Fuzzy Type 2 have a ability to fast responses for recovery on muscle fatigue and muscle force potential. This fast responses very needed for clinically actions for swing phase gait restoration on standing subject.

In the future, we would make experiments muscles contraction restoration for stroke patient in clinically. We still completely safety utilizing for clinically used, safety for patient, and also safety for medical officer like physicians or nurses.

5. ACKNOWLEDGMENTS

Our thanks to Achmad Arifin as my supervisor for my thesis on Sepuluh Nopember Institute of Technology. Also to another students who want to be a subject for my research.

6. REFERENCES

[1] Chaoui, H., Gueaieb, W. 2007. Type-2 Fuzzy Logic Control of a Flexible-Joint Manipulator. J Intell Robot Syst. 51 (2008), 159-186. DOI= 10.1007/s10846-007-9185-2

[2] Arifin, A. 2005. A Computer Simulation Study on the Cycle-to-Cycle Control Method of Hemiplegic Gait Induced by Functional Electrical Stimulation. A Doctoral Dissertation, Japan: Tohoku University.

[3] Karnik, N.N., Mendel, J.M. 1998. Introduction to Type-2 Fuzzy Logic Systems. IEEE Fuzz Conf.

[4] Wicaksono, H. 2009. Fuzzy Controller Type 2 Based on Cycle-to-Cycle Method for Swing Phase Gait Restoration with Functional Electrical Stimulation. Thesis. Sepuluh Nopember Institute of Technology. Surabaya, Indonesia.

[5] Wicaksono, H. 2009. Swing Gait Phase Restoration using Fuzzy Controller Type 2. SNTF Conference. Surabaya, Indonesia.

Design and Construction of Wind Speed Indicator Based on PIC Microcontroller System

Khin Mar Aye
Computer Hardware Department
Computer University, Monywa
Myanmar
kmayester@gmail.com

Kyi Thar Oo
Computer Hardware Department
Computer University, Monywa
Myanmar
kyitharoo10@gmail.com

ABSTRACT
Wind speed is measured in a wide variety of ways, ranging from simple to most sophisticated electronic systems. This wind speed indicator is intended for use in a variety of activities, such as track events, sailing, hand-gliding, model aircraft flying, and measuring the speed of airflow in wind tunnels or in other gas-flow application. In this wind speed indicator, two ultrasonic transducers and PIC microcontroller are used. Knowing the basic speed of sound under specified conditions, the rate at which the air mass moving can be calculated from the measured timing. The read out is shown on an alphanumeric liquid crystal display (LCD), with reading in meter per second, feet per second, kilometer per hour, and mile per hour. In this constructed anemometer, only the current wind speed can be measured and the average wind speed taken over 16 transmission cycles can also be obtained. There are two ways of calibrations. Firstly, by comparing the wind speed of constructed anemometer with the car's speedometer, it is accurate up to 50 kilometer per hour, and then began to fall of rapidly. Secondly, the speed of constructed anemometer is compared with reference anemometer. By comparing the resulting data, the magnitude of wind speeds are nearly the same. The resolution of constructed anemometer is to the nearest tenth of a meter per second and the measuring range of wind speed is from zero up to 50 miles per hour and possibly higher.

Keywords
Ultrasonic transducer, PIC microcontroller, Anemometer.

1. INTRODUCTION
The instrument that measure wind speed is called anemometer. Several techniques for measuring wind speed exist. Wind speed is measured in a wide variety of ways, ranging from simple to most sophisticated electronic systems. The cup type anemometer has three or six light, conical cups mounted on a shaft which turns on low-friction bearings [1]. The cup type anemometer was constructed by Ni Ni Khaing, at Yangon Technological University, in 2003 [5]. The sonic anemometer is the most suitable type for advanced wind measurements, particularly where it is necessary to capture very fast fluctuations. Design of 1D sonic anemometer was constructed by the Department of Electrical and Computer Engineering, at University of Massachusetts Amherst, in 2003 [7]. An acoustic digital anemometer was also constructed by Tufan coskun Karalar, Department of Electrical Engineering and Computer Science, University of California at Berkely [6]. The goal of this anemometer is to design and implement a system that can be used for monitoring indoor airflow.

This ultrasonic wind speed measurement technology uses ultrasound to determine horizontal wind speed. The ultrasonic technique measures the transit time, the time it takes for the ultrasound to travel from one transducer to another, depending on the wind speed along the ultrasonic path. With wind along the sound path, the upwind transit time increases and the downwind transit time decreases. From difference transit times, the PIC (Peripheral Interface Controller) microcontroller computes the horizontal wind speed. The rest of paper is organized as follows.

2. AIM AND OBJECTIVES
This ultrasonic wind speed indicator (anemometer) is intended for use in a variety of activities, such as track events, sailing, hand-gliding, and model aircraft flying, to name but a few. This is also used instrumental in measuring the speed of airflow in wind tunnels or in other gas-flow application. It can also be used to monitor the conditions in the garden or compound. The objectives of this research are as follows:

- To share the basic principle of acoustic digital anemometry
- To describe the ultrasound and the operation of ultrasonic transducer system
- To design one dimensional anemometer that will measure small wind speed
- To send and receive ultrasonic signals across transducers that are not affected by noise

3. BASIC PRINCIPAL OF ACOUSTIC ANEMOMETER
The main fact taken by acoustic digital anemometry is that sound propagation speed is directly affected by the motion of its propagation medium. For sound wave traveling in air, any air flow affects the propagation speed. This is shown in figure 1.

For example, there are two points 1 and 2 and there is also airflow from point 1 to point 2 along the line connecting these two points. If the sound of speed is measured from point 1 to point 2, the result will be

$$V_{12} = V_{sound\text{-}still} + V_{airflow} \quad (3.1)$$

Where V_{12} is the speed of sound with point 1 as the transmitter and point 2 as the receiver, $V_{sound\text{-}still}$ is the speed of sound in still air and $V_{airflow}$ is the airflow speed along the line from point 1 to point 2.

Then, if an acoustic signal is transmitted in the opposite direction, i.e., point 2 is the transmitter and point 1 is the receiver, the sound speed will be measured as

$$V_{21} = V_{sound\text{-}still} - V_{airflow} \quad (3.2)$$

where the variables V_{21}, $V_{sound\text{-}still}$, and $V_{airflow}$ are defined similarly to those in equation (2.1).

If V_{12} and V_{21} are measured for these two cases, $V_{airflow}$ can be obtained using

$$V_{airflow} = (V_{12} - V_{21})/2 \quad (3.3)$$

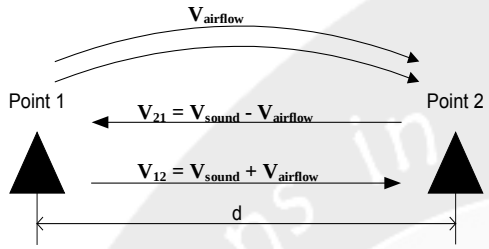


Figure 1. Basic Principle of Airflow Measurement

The directionality of the flow can be easily determined by nothing that airflow has the same direction as the transmission yields a higher acoustic speed. It should, however, be kept in mind that by taking the difference of two speeds the result would be only the component of airflow along the line connecting point 1 and 2.

As already discussed, speed of sound is a strong function of the properties of the propagation media, i.e. temperature, type, state, etc. This dependence can significantly affect the $V_{sound\text{-}still}$ components. But in the difference of V_{12} and V_{21}, $V_{sound\text{-}still}$ term drops out and this strong dependence on ambient condition is not a main concern .

4. MICROCONTROLLER

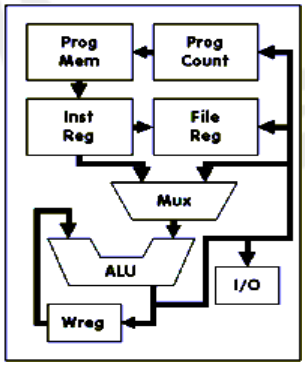


Figure 2. A Block Diagram of PIC Microcontroller System

A microcontroller is a single chip computer Micro suggests that the device is small, and controller suggests that the device can be used in control applications. Another term used for microcontroller is embedded controller, since most of the microcontrollers are built into the devices they control. It is one of the most important developments in electronics since the invention of the microprocessor itself. All microcontrollers operate on a set of instructions stored in their memory.

PIC has the calculation function and the memory like the CPU and is controlled by the software. However, the throughput, the memory capacity isn't big. It depends on the kind of PIC but the maximum operation clock frequency is about 20 MHz and the memory capacity to write the program is about 1K to 4K words.

In this research, PIC16F628-20/P was used to generate the 40 kHz pulse and to control the route of pulse between the ultrasonic transducers, and then to calculate the wind speed.

The PIC 16F628A belongs to midrange family of the PIC microcontroller devices. It has the following features [6]:

1. 8K word of program memory
2. 68 bytes of RAM
3. 128 bytes of EEPROM
4. 16 input / output pins
5. timer and interrupt functions
6. comparator function

5. DESIGN IMPLEMENTATION OF WIND SPEED ANEMOMETER

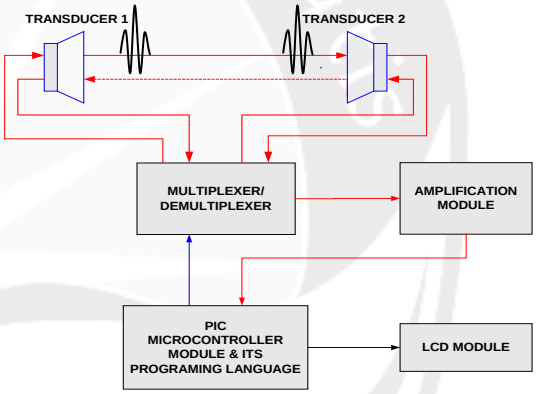


Figure 3. Block Diagram of Constructed Wind Speed Meter

A proposed design for a one dimensional ultrasonic wind speed meter is described in figure 3. Sound waves can be used as a tracer to measure wind speed. The sound is injected into the atmosphere, and the time taken for it to be carried over a fixed distance is measured. The anemometer measures the difference in transit times between sound pulses moving between ultrasonic transducers.

The time it takes for a sound to travel between a source and receiver can be easily measured. Knowing the basic speed of sound under specified conditions, the rate at which the air mass is moving can be calculated from the measured timing. When using a single source and receiver, of course, the wind must be moving directly in line with them. In practice, it does not matter whether the wind flows toward or away from the source, electronic techniques can compensate accordingly.

The use of an audio sound source and receiver would not be practical since such a system would be subjected to interference from many extraneous sounds. Ultrasonic method, though, are much less susceptible to interference. Using the ultrasonic transducers, the wind speed assessment is easy to understand. Two ultrasonic transducers are faced each other across a known

distance. One shoots a pulse at the other and the time it takes for the signal to cross between the two is measured.

5.1. Ultrasonic Sensing and Obtaining Wind Speed

5.1.1. Ultrasonic Signal Transmission and Receiving

The circuit diagram for the ultrasonic transmission and reception functions is shown in Figure 4. The two transducers are shown as X3 and X4. As just said, they are both used interchangeably as transmitter and receiver. Analogue multiplexer IC3 selects the mode in which the transducers are used.

The transducers operate at usual ultrasonic frequency of 40 kHz. The transmission pulses are generated by a PIC microcontroller, which is described presently in relation to Figure 4.2. The route that the pulses take through IC3 is selected by the logic level applied to its pin 10, also controlled by the PIC.

When pin 10 is held low, the pulses are routed from IC3 pin 3 to pin 1, and out to transducer X3. This transducer transmits the pulses across a gap of several centimeters to the second

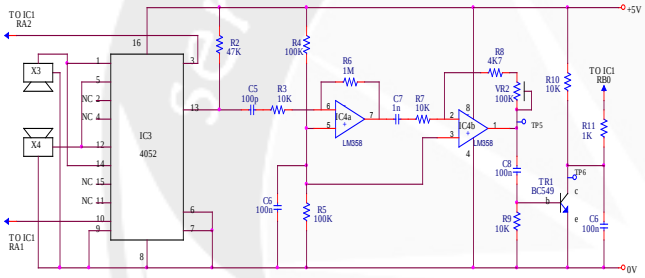


Figure 4. Ultrasonic Transmissions and Reception Circuit Diagram for the Ultrasonic Wind Speed Meter

transducer, X4, which receives the pulses and routes them to IC3 to pin 12. The pulses pass through IC3 to pin 13 and to the analogue amplification circuit formed.

When IC3 pin 10 is held high, the pulses are routed from IC3 pin 3 to pin 5, and this time out to transducer X4. Now transducer X3 receives them and they pass via pin 14 to pin 13 and so out to the amplifier.

5.1.2. Received Pulse Signal Amplification

The pulses pass through IC3 to pin 13 are much attenuated by their journey and the analogue amplification circuit is formed around op-amps IC4a and IC4b. This is shown in Figure 5. In this design LM358 op-amps are used as ac coupled inverting amplifiers. The ac-coupled inverting amplifier multiplies the non-DC input voltage by the desired negative gain:

$$V_{out} = [-(R_f / R_g) \times V_{in}] + V_{bias} \quad (5.1)$$

where V_{bias} is added to the op amp non-inverting input to bring that input pin voltage within the normal operating range of the amplifier. It also provides an offset for the output, so that it is within its operating range. The closed loop gain of inverting amplifier is

$$A_c = -\frac{R_f}{R_i} \quad (5.2)$$

From IC3 pin 13, the received pulses are ac coupled via capacitor C5 to the first amplifier, IC4a. The gain of first amplification stage is

$$A_{c1} = \frac{R_6}{R_3} = \frac{1M}{10K} = 100$$

A gain of about 100 is provided by this stage, as set by the values of resistors R3 and R6. The signal is then ac coupled by C7 to the stage around IC4b.

$$A_{c2} = \frac{R_8 + VR_2}{R_7} = \frac{4.7K + 0K}{10K} to \frac{4.7K + 100K}{10K} \cong 0.5 to 10$$

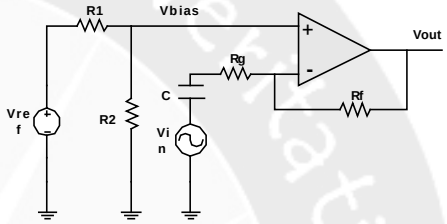


Figure 5. The Received Ultrasonic Pulse Amplification Circuit

Here the gain can be varied between about x0.5 and x10, as controlled by preset VR2. The potential divider formed by R4 and R5 applies mid-rail bias to the non-inverting inputs of the two op-amps (pins 5 and 3 respectively).

The final gain stage is provided by transistor TR1. Its base (b) is biased normally low by resistor R9, so holding it in a turned-off condition. The output from IC4b is ac coupled to TR1 by capacitor C8 which exceed about 0.6V turns on TR1, causing a full line-level negative-going pulse at its collector(c). This pulse is coupled via resistor R11 back to the PIC.

5.1.3. Control Circuit and Output Display

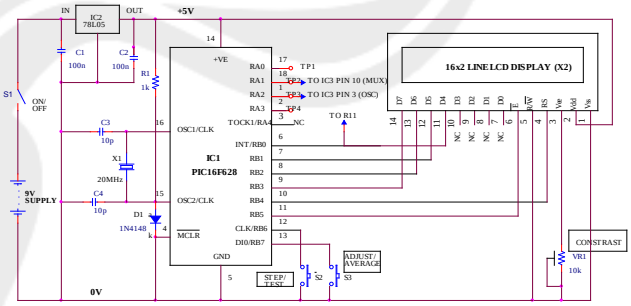


Figure 6. Circuit Diagram for the Control and Display Functions of Ultrasonic Wind Speed Meter

As shown in the control circuit diagram of Figure 6, the PIC 16F628 microcontroller (IC1) is responsible for generating and sending pulses to the ultrasonic transducers, and for timing the return of the received signal. The PIC is operated at 20MHz as set by crystal X1 in conjunction with capacitors C3 and C4.

The results of calculations are output to the 2-line 16-character alphanumeric LCD(X2). This is operated in 4-bit control mode, with its screen contrast adjustable by preset VR1.

5.2. Software Implementation

As shown in the control circuit diagram of Figure 6, the PIC16F628 microcontroller is responsible for generating and sending pulses to the ultrasonic transducers, and for timing the return of the received signal. The wind speed is calculated from the difference timing value between two timing values. The results of its calculations are output to the 2-line 16-character alphanumeric LCD display.

The process for the transmission and receiving of ultrasonic signal, calculation of wind speed and display of these values are written with MPASM assembly language.

5.3. Ultrasonic Transducer Assembly

There are two types of ultrasonic transducer assembly in this design, one for hand held and another for settle mount, shown in Figure 7. For hand-held, two ultrasonic transducers are mounted on the aluminium stick and faced each other. They are distant 18 cm (7 inches) and the aluminium stick is one foot long. For settle mount**,** the ultrasonic transducers are mounted on the shaft of the wind vane. The distance between the transducer faces was set to about 18 cm (7 inches), but the distance is not critical and a fraction either way does not matter.

The transducers used in this design were the standard front-facing open-mesh type. It does not matter in which order the transducers are mounted and connected. Although supplied as a pair comprising one transmitter and one receiver, as explained earlier, they are used interchangeably in both capacities.

(a) **(b)**

Figure 7. Transducers Assembly of Ultrasonic Wind Speed Meter: (a) Hand- held (b) Settle-mount.

5.4. Display Values

When the power is switched on, four sets of values will be seen on the LCD, possibly changing a bit erratically at present. This is shown in Figure 8(a). On the top line are shown the monitored wind speed values in meters and feet per second, both having two decimal places to the nearest 0.01 value. The maximum integer value that can be shown is 99.

The lower line shows the speed in kilometer and mile per hour, to one decimal place, with a maximum integer value of 999. In fact, it is not actually known how high a wind speed the unit will correctly respond to, but it should be at least 50mph (80kph) and likely to be much higher.

(a) **(b)**

(c)

Figure 8. Wind Speed Display Values: (a) for Current Value (b) for AverageValue, (c) for Timing Value during Each Pair of Transmission Cycles

When the switch S3 is pressed, the unit into full averaging mode is set and signified by the letter 'Av' being shown at the far right of LCD line 2, shown in Figure 8(b). In this mode, the second block of 16 values previously mentioned is averaged and the calculations use that result instead of the immediate value that is shown when averaging is off, and 'Av ,' replaced by two blanks on screen. Repeated pressing of S3 toggles between the two modes.

When the switch S2 is pressed, the test mode is selected by replacing the top line values with the actual timing values, detected during each pair of transmission cycles, shown in Figure 8(c). These are the actual values read from the PIC's Timer 1 register. To their right is shown the absolute difference between them (without + or – signs).

When the switch S2 is pressed again, once more it is caused the meters per second (m/s) and feet per second (f/s) speeds to be shown.

6. Test and Results

It is necessary to calibrate in every constructed anemometers. There are two ways of calibration. The constructed anemometer is calibrated by placing it in the known wind speed airflow of wind tunnel. However, it is not possible because it is not obtained wind tunnel. Thus, the constructed anemometer is calibrated by adjusting the resulting values of constructed anemometer with the available commercial reference anemometers. Before it is not adjusted with reference anemometer, the wind speed of constructed anemometer is compared with the car's speedometer. The ultrasonic transducer assembly is positioned outside the car's window in this measuring. The measuring of constructed anemometer is accurate up to about 50 kilometer per hour, and then began to fall of rapidly. It is concluded that the aerodynamic of the car began to take effect above this speed. The comparison of speed of car and speed of ultrasonic anemometer is shown in Figure 8.

Then this constructed anemometer is compared with reference anemometer. The reference anemometer is Sensitive Anemometer (1186 – Casella, London, England) that reads in feet per second. In this data reading, the reference anemometer reads the wind speed

at 30 second average values. The constructed anemometer reads the wind speed at 5 second average value in 6 times. Then, 30 second average value is calculated. By this ways, the data reading of the two anemometers is obtained.

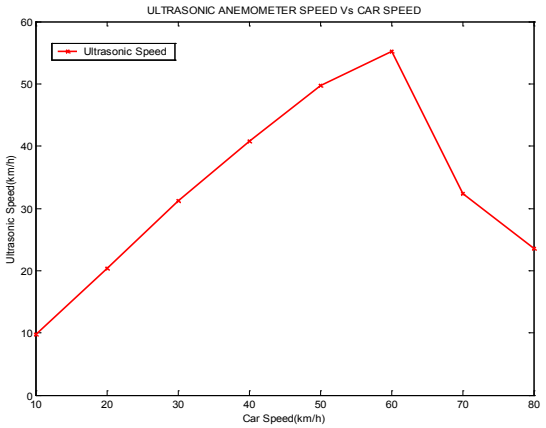


Figure 9. A Graph of Comparison of Car's Speed and Ultrasonic Anemometer

By comparing the resulting data, the magnitude of the recorded values from the reference anemometer and constructed anemometer are nearly the same. The result of the constructed anemometer is 4.4034% less than the result of reference anemometer. The comparing graph of resulting data of these anemometers is shown in Figure 9.

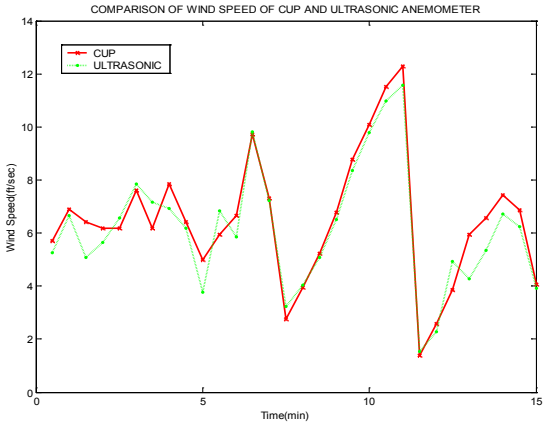


Figure 10. A Graph of Comparison of Reference Anemometer and Constructed Ultrasonic Anemometer in ft/s

7. CONCLUSION

The ultrasonic transducer assembly is pointed in the direction from which the wind is blowing and a screen displays the rate at which the wind is moving between two ultrasonic sensors. The readout is shown on an alphanumeric display (LCD), with reading in meters per second, feet per second, kilo-meters per hour, and miles per hour. In this constructed anemometer, only the current wind speed can be measured and the average wind speed taken over 16 transmission cycles can also be obtained. The resolution is to the nearest tenth of a meter per second and the measuring range of wind speed is from zero up to 50 mile per hour and possibly higher.

The advantages of this anemometer are:

(1) The threshold is zero. Therefore, it is more sensitive than others.

(2) It can measure very small wind speeds accurately.

(3) Since it is measuring time of flight, the output is linear.

(4) It can measure wind speed three times per second.

(5) It can measure from zero up to 50mile/hour, and possibly higher.

(6) Resolution is 0.1 m/sec.

8. LIMITATION

The anemometers are the instruments that measure wind speed. Generally two principal components of the wind, such as wind speed and direction, are most interested parameters. Wind direction is also an important item of information.

Although this constructed anemometer is fitted on the wind vane to detect the wind direction, it does not display the directions of the wind in the display system. It is used just to mention the direction of the wind because the aim of this research is only to measure the wind speed. In this design, the recording system will not be obtained because the constructed anemometer is intended for to indicate the current wind speed. However, if it is necessary to display the wind direction and to record the data, some modification has to be made.

9. REFERENCES

[1] G.L. Johnson, "Wind Energy Systems", New Jersey: Prentice-Hall, 1985.

[2] G.S. Campbell, "Biophysical Measurements and Instrumentations", 1996.

[3] G.S. Campbell & M.H. Unsworth, "*An Inexpensive Sonic Anemometer for Eddy Correlation*", J. Appl. Meteorol, Vol.18, No.1027, 1979.

[4] M. James, "Microcontroller Cook Book PIC & 8051", 2nd Edition, Newnes, Butterworth-Heinemann, 2001.

[5] N.N. Khaing, "Design and Construction of Microcontroller Based Wind Speed Recorder System", Yangon Technology University, 2003.

[6] T.C. Karalar, "An Acoustic Digital Anemometer", Department of Electrical Engineering and Computer Science, University of California at Berkeley.

[7] V. Dube, M. Jao, C. Srinivasa & R. Vice, "*Design of One-Dimensional Sonic Anemometer*", Department Electrical and Computer Engineering, Massachusetts Amherst University, December ,1999.

[8] "PIC16F62X (FLASH-Based 8-Bit CMOS Microcontrollers)", ©Microchip Technology, 1999. http://www.microchip.com.

Fault Diagnosis in Batch Chemical Process Control System Using Intelligent System

Syahril Ardi
Faculty Member, Politeknik Manufaktur Astra
Jl. Gaya Motor Raya No. 8, Sunter II – Jakarta 14330
62-21-6519555
syahril.ardi@polman.astra.ac.id

ABSTRACT
Batch chemical industries have been attracting for safety engineers since they pose a number problem in behavior and operation. The reliability of a chemical reactor installed in a plant depends on the capability of the control/ supervision system to estimate its state and to identify, in time, its operational malfunctions or failure modes. In this paper, fault diagnosis in batch chemical process control system using intelligent system is proposed. The artificial intelligence technologies that are associated with knowledge-based approaches and adopted for monitoring, control, and diagnosis in the process industries include expert systems, fuzzy logic, neural network, and support vector machine.
As has been mentioned, a correct choice of reactor operating conditions does not totally protect the plant against a thermal runaway. So, apart from the off-line activities, which help to define safe operating conditions, also on-line prevention measures are necessary to detect any unexpected situation leading to a runaway scenario. Among others under the on-line safety measures, an early warning detection system is indispensable to detect and evaluate unexpected dangerous situations, which may occur in batch reactors e.g., due to a failure of the cooling or stirring systems or to a human mistake. Nowadays, the batch industries are seeking for the more real time, accurate, efficient and low-cost method and application for supporting safety in their industries. The use of intelligent system method (comparing between neural network and support vector machine) for fault diagnosis in chemical batch plant can be the best choice for the solutions. The solutions, which are recommended by the use of intelligent system, will support the operator in their activities to control, prevent, and mitigate the hazards.

Keywords
Batch chemical process control, neural network, support vector machine, fault diagnosis

1. INTRODUCTION
Batch chemical industries have been attracting for safety engineers since they pose a number problem in behavior and operation. Compare to the continuous plants, a batch process has unique characteristics, behavior of process changes dynamically, role of operators, and the change of process variable [1]. Besides using the hazardous material, a batch abnormality can be caused by the deviation of process variable and plant mal-operation [2]. A process variable deviation occurs during batch process and it becomes a significant factor for the safety issues in the plants. The deviations tend to influence plant operation and change the situation into abnormal state and contribute in damaging the plants.
Towards fulfill the global requirement, the companies should ensure safety and quality for their plant. Over the years, product quality has become primary focus due to it gives direct impact to benefit. However, due to aging and reducing of plant reliability, safety problems will emerge in some existing batch plants. Some conventional methods such as Failure Tree Analysis, Event Tree Analysis, and Human Reliability Analysis are still used in industries. However, these method are capable only both in pre-accident and post-accident situation [3], and do not consider the real condition of the plants, it is difficult to analyze accurately the real time process and integrate the results with the prior condition of the processes. For example, Event Tree Analysis, it is usually implemented in two stages, pre-incident and post-incident analysis. Pre-incident is intended to identify hazards that vulnerable to the plant system, environment, human, etc., the results should be considered as the input for operational stage. The post-incident analysis is intended to evaluate the probability of occurrence, the results of this analysis should be considered as historical data that guides the operational stages for the near future. On the other hand, identification of hazards based on the operational stage condition is essential to avoid the consequences and revise the recent condition for safety improvement program. As solution of this problem, a fault propagation assessment technique can be used in analyzing accurately the abnormalities. Fault propagation manipulates information from sensor and batch recipe system to estimate the performance of safety objects in handling abnormality [4]. The aim of real time hazard identification is to predict the condition patterns prior to accident occurs, these patterns are expressed in indices and obviously, the optimal policy for plant safety design and maintenance can be decided in avoiding the hazardous outcomes. The real-time applications of performance reliability prediction are useful in operation control as well as predictive maintenance [5].
One of the potential accidents in the batch plants is runaway reaction. Runaway reaction has been reported as potential hazard in the batch process. Based on the newest data, runaway reaction leads to fires and explosion, where the majority of incidents occurred during normal operation and two main causes were investigated as runaway reaction and overflow of material. As case study, runaway incidents in PVC batch process are investigated at VCM charging line that potential contributes to overflow of monomer and lead to runaway reaction.
The objective of this paper is to evaluate the application of Support Vector Machine (SVM) for predicting runaway reaction in a PVC reactor. This research is a continuation of our research before [6]. SVM models, which are based on the statistical

learning theory, are a new class of models that can be used for predicting the values.

2. FAULT DIAGNOSIS IN BATCH PROCESSES

Most of the fault diagnosis approaches presented so far are applicable to steady-state processes. These approaches can be divided into three groups: historical based methods, model-based techniques and combinations of both. However, the application of these diagnosis approaches to batch chemical processes is usually difficult.

In the past decade, research was focused on the use of either fundamental models or detailed knowledge based models. The first monitoring procedure is based on estimation methods. The second relies on the knowledge of the operators and engineers about the process.

More recently, the use of pattern recognition methods based on neural networks and the use of statistical techniques are matter of research. With respect to the use of statistical techniques, multiway principal component analysis (MPCA) has shown good results in batch process monitoring (Nomikos and MacGregor, 1994). This technique is currently used as a reference in the present research. The only information needed is a historical database of past batches. However, it has some drawbacks like the difficult isolation and localization of the fault.

Finally, in order to combine the strengths of both pattern recognition and inference methods, adaptive neuro-fuzzy systems are being developed. The idea is to obtain an adaptive learning diagnostic system with transparent knowledge representation. Some combinations are the subjects of current research (Leonhardt and Ayoubi, 1997): ANNs influenced by fuzzy logic (e.g. fuzzy models within ANNs), fuzzy systems influenced by ANNs (e.g. serial configuration), and hybrid neurofuzzy systems. In the past few years, the application of combined methods for fault diagnosis has steadily been growing.

3. ANNs FOR FAULT DIAGNOSIS WITH CASE STUDY IS PVC BATCH PROCESS

Among the pattern recognition methods, the ANNs approach is the most popular. ANNs have many very useful properties for fault diagnosis. They can handle nonlinear and undetermined processes. They learn the diagnosis by means of the training data. They are very noise tolerant and work well with noisy measurements.

Their ability to adapt during use is also an interesting property. In the petrochemical industry, ANNs have been used as supervised pattern classifiers. They are trained on historical or simulated steady state process data with the aim of detecting a specified number of suspected faults.

The first reports (Hoskins & Himmelblau, 1988; Venkatasubramanian, Vaidyanathan & Yamamoto, 1990) show the application of backpropagation networks (BPNs) using sigmoidal functions in the first layer. In more recent studies, radial basis function networks (RBFNs) are preferred because they provide more reliable generalisation and fewer extrapolation errors (Yu, Gomm & Shen, 1998). The elliptical basis function neural network is similar to RBFNs with Gaussian basis function. However, it has more favorable and intuitive results in function approximation and classification (Chen & Liu, 1999). Self-organizing maps (SOMs), which are trained, unsupervised, are not always able to classify data correctly.

However, their ability to classify data autonomously is very interesting and useful when real industrial processes are considered (Koivo, 1994). Regarding the special case of faults in sensors, auto-associative neural networks have been showing good results. Their application is based on the nonlinear principal component analysis technique (Kramer, 1992).

The problem of the traditional ANNs related to totally capture the space and time characteristics of process signals are overcome with the use of wavelet functions. Studies on wavelet functions, an area of signal processing, have advanced rapidly in the last few years. Its application to fault diagnosis is being performed in two ways:

- By using it for feature extraction, the outputs are then processed either by an ANN (Chen, Wang, Yang & McGreavy, 1999), by qualitative trend analysis (Vedam & Venkatasubramanian, 1997), or by a principal component analysis approach (Bakshi, 1998).
- By using it as an activation function in an ANN (Zhao, Chen & Shen, 1998).

3.1 Detection and Identification of Runaway Application

The runaway reaction is generally accepted to occur in stages, *viz.*,

Detection – to ascertain, if a runaway has occurred?

Identification – where exactly is the fault?

For this case study, I have selected a model plant of PVC batch, as shown in Fig. 1. In the batch reactor of PVC plant, the raw materials including the reactants and catalysts are charged to the reactor first and agitated, after that the reaction is initiated with heat being added or removed as required. From this simplify model plant, a batch dynamic simulator using Visual Modeler (VM) has been developed and from this we obtained the process variable data on temperature, pressure, and level of the reactor (Rizal et al., 2005). Next, this data is used for detection and identification of process fault in case of a runaway reaction. The focus of this chapter is on recognizing condition of the temperature normal and abnormal (runaway reaction) inside the reactor. The experimental results of the temperatures inside the reactor, for normal and abnormal reactions are shown in Fig. 2. In Figure 2, the temperature inside the reactor has been divided into three zones, *i.e.*, charging, heat-up, and reaction tasks. During the reaction step, the temperature is maintained between 50 ^{0}C and 60^{0}C by charging the cooling water. We have observed that the temperature inside the reactor will start deviating when the process transits from heat-up phase to reaction phase. This temperature deviation shows the possibility of a runaway reaction.

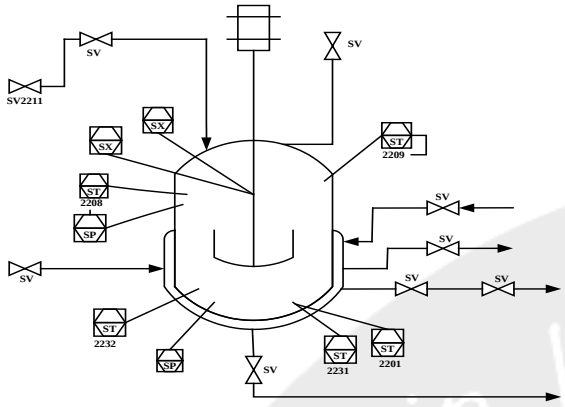


Figure 1. Batch reactor of PVC plant diagram

Reactor model must contain all the information pertaining to detection; first, if the batch operation is going beyond the normal limits of temperature and pressure. If an abnormal situation is encountered *(i.e.*, runaway reaction), the system must identify the fault.

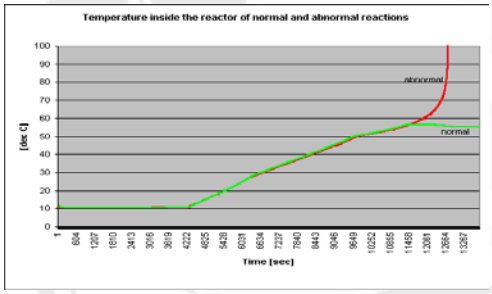


Figure 2. The experiment results of the temperature inside the reactor: normal and runaway reactions

4. SUPPORT VECTOR MACHINE

Various machine-learning techniques, for example neural networks, have already been widely used in computational chemistry. In the last years, however, neural networks have been used somewhat less in engineering and science. Instead there has been an increasing interest in support vector machine for various domains, due to its ease in handling complex problems.

Support vector machine is a novel statistical learning theory based on machine learning algorithm presented by Vapnik and coworkers in 1995. Originally, SVM was developed for solving classification problems. Recently, with the introduction of ε-insensitive loss function, it has been extended to solve regression problems, and has shown great performance in QSPR (Quantitative structure–property relationship) studies due to its ability to interpret the nonlinear relationships between molecular structure and properties [7].

4.1 Model Implementation With Case Study – PVC Batch Plant

Case study is simplifying batch reactor of PVC plant as has shown in Figure 1. The temperature data (normal and runaway) are used to illustrate the performance of the proposed model. This study describes a classification methodology based on SVM, which using SVM classifier with linear kernel function and one-norm function.

Software program that utilized SVM model is the Matlab with Bioinformatics toolbox (The Mathwork, 2009). Steps for classify data using support vector machine are:

- Load the data (temperature of vessel)
- Create *data*, a two-column matrix containing sepal length and sepal width measurements
- From the *species* vector, create a new column vector, groups, to classify data into two groups: Normal and Abnormal
- Randomly select training and test sets
- Use the *svmtrain* function to train an SVM classifier using a linear kernel function and plot the grouped data.

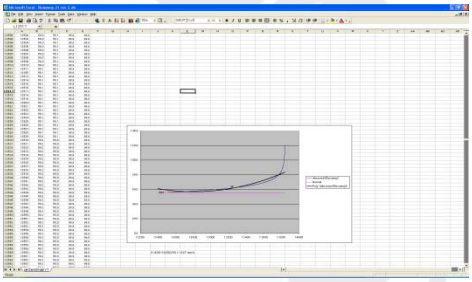


Figure 3. Data's temperature in reactor PVC batch plant

In this implementation, I have used data's temperature in reactor PVC batch plant as shown in Figure 3. The data's temperatures consist of normal and runaway condition.

Figure 4 represent the simulation result of SVM classify the test set using a support vector machine from temperature data in reactor PVC batch plant. From this figure have been found an optimal separating hyperplane by maximizing the margin between the separating hyperplane and the data. In this case, the y- axis is for normal temperature and the x-axis is for the runaway temperature (in degree Celsius).

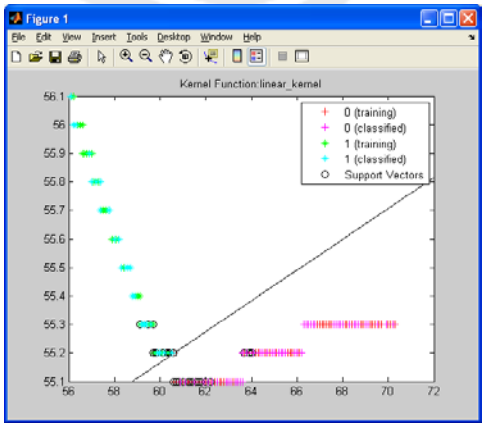


Figure 4. Simulation result of SVM classify the test set using a support vector machine from temperature data in reactor PVC batch plant

After that, we can also make an evaluation the performance of the classifier with formula and result (Matlab execution):

```
>> classperf(cp,classes,test);
>> cp.CorrectRate
ans =
   0.9867
```

Figure 5 and Figure 6 represent the simulation result of hard margin SVM classifier by using a one-norm. For evaluating the performance of the classifier, we use the formula and result (Matlab execution):

```
>> classperf(cp,classes,test);
>> cp.CorrectRate
ans =
   0.9867
```

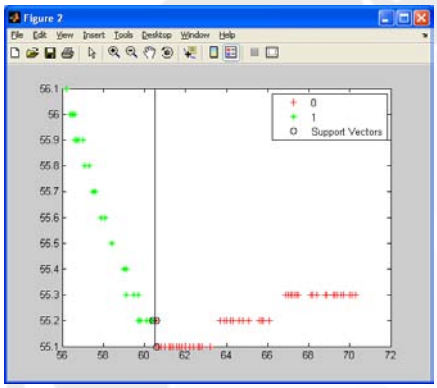


Figure 5. Simulation result of SVM hard margin SVM classifier (svmtrain) by a using one-norm

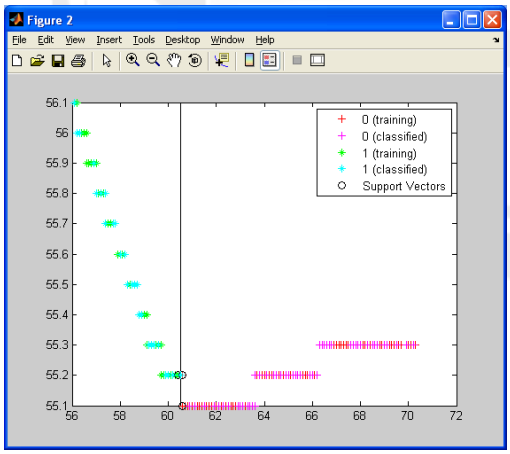


Figure 6. Simulation result of SVM hard margin SVM classifier (svmclassify) by using a one-norm

From simulation result of SVM classify the test set using a support vector machine from temperature data in reactor PVC batch plant, we have got that the optimal hyperplane is F(x) = 60.6 ^{0}C. From the data, we note that the temperature 60.6 ^{0}C occurred when the time process is 13105 s. If we assume that runaway reaction will start when the temperature threshold value is 70 ^{0}C (13614 s), then we have time for anticipating the runaway reaction is about 8.5 minutes.

5. CONCLUSIONS

Study of fault diagnosis in chemical batch plant using intelligent system (knowledge-base) is described and analyzed in this work. In the context of chemical batch plant where hazardous materials are used, it is highly critical to monitor the activities of process to prevent the accident from occurring. Obviously, inconsistencies of process variables during operational stage often occurred. In addition, the plant miss-operations caused by the plant operators also contribute to the harmful situation. In order to overcome the chemical batch problems, the applications of fault diagnosis using intelligent system are presented. Employing different approaches for fault diagnosis such as model-based that uses quantitative models and equations to estimate the states or parameters of the system. The general conclusions from these studies are:

Applying neural networks to detect runaway reaction in a batch process, we have mentioned the parameters *m*, *b*, and *R* that measure the efficiency of trained network, for normal and runaway conditions. For investigating the causes leading to undesired consequence (abnormal condition) and where this occurs, we have used fault tree analysis.

Study on predicting runaway reaction using support vector machine (SVM) is considered for their simpler design & implementation, and for allowing the better handling of complex & large data sets. From the simulation result of SVM classifier from temperature data in reactor PVC batch plant, we found that the prediction of runaway reaction can be detected earlier and safer i.e., we have time for anticipating the runaway reaction is about 8.5 minutes.

6. REFERENCES

[1] Aamodt A, Plaza E., CBR:Foundation issues, methodological variations, and system approaches, Artiff Intell Commun, Vol. 7(1), pp.39-59, 1994

[2] ANSI/ISA-S88.01-1995.(1995).*Batch control part 1: models and terminology", Instrument Society of America* (1995)

[3] Ardi, S., Datu Rizal, Shinichi Tani, Kazuhiko Suzuki and Hossam Gabbar, "Detection and Identification of Runaway Reaction in a Batch Process Using Artificial Neural Network", The World Conference on Safety of Oil and Gas Industry, TS VI-6, pp.378-381, Gyeoungju, Korea, 2007.

[4] Ardi, S., Hirotsugu Minowa, Kazuhiko Suzuki, "Failure Detection Algorithm Based on Intelligent Monitoring of Safety Components, Proceeding of the 2008 International Joint Conference in Engineering, IJSE2008, August 4-5, Jakarta, Indonesia.

[5] Ardi, S., Hirotsugu Minowa, Kazuhiko Suzuki, "Detection of Runaway Reaction in a Polyvinyl Chloride Batch Process Using Artificial Neural Networks", International Journal of Performability Engineering, Volume 5 – Number 4, pp. 367-376, July 2009.

[6] Ardi, S., Minowa, H., Suzuki, K. (2008). Failure Detection algorithm based on intelligent monitoring of safety components, *Proceeding of the 2008 International Joint Conference in Engineering IJSE2008*, August 4-5, Jakarta, Indonesia, 137-140.

[7] Yong Pan, et al., Predicting the auto-ignition temperatures of organic compounds from molecular structure using support vector machine, Journal of Hazardous Materials 164 (2009) 1242–1249.

Implementation of An Adaptive PID Controller Using The SPSA Algorithm with Realistic Target Response

Sofyan Tan
Universitas Pelita Harapan
Karawaci, Tangerang
sofyan.tan@gmail.com

ABSTRACT
An adaptive control can be very useful in applications where the characteristic of the controlled plant is changing over time. This paper describes one implementation of an adaptive proportional-integral-derivative (PID) controller using the simultaneous perturbation stochastic approximation algorithm (SPSA). The SPSA algorithm is appropriate for this application because it is usually difficult to obtain the detailed nonlinear model of the system response with respect to the PID gains, and the number of measurements required is relatively small. The proposed adaptive controller is a model reference adaptive control using a realistic target response instead of an ideal step function. The evaluation showed that the proposed adaptive controller is able to optimize the PID gains with the output position response approaching the target position response.

Keywords
Adaptive control, PID controller, SPSA algorithm

1. INTRODUCTION
The traditional proportional-integral-derivative (PID) controller holds an important role in many applications where close loop control systems are required. In many implementations, the PID gains of the controller need to be calibrated manually to meet certain objectives. However, over a time period, the characteristic of the plant usually changed, and must be recalibrated periodically to meet the objectives. In such applications, where the plant or the environment significantly changes over time, an adaptive controller can be very useful since it can automatically adapt the PID gains to achieve the desired objectives.

This paper evaluates the implementation of an adaptive PID controller using the simultaneous perturbation stochastic approximation (SPSA) algorithm [1-2] and a realistic target response modeled as a linear time invariant system. The SPSA optimization algorithm is suitable for multivariate nonlinear problem where the model of the controlled plant is difficult or impossible to obtain.

Adaptive PID controller using the SPSA algorithm has been observed in [3], where the target response was an ideal step function. This paper likes to investigate the performance of the adaptive controller where the target response is a more realistic function. Furthermore SPSA algorithm in this paper is slightly modified from [1-2] to include different gain sequences for different PID gain.

2. THE ADAPTIVE CONTROL METHOD
Generally the adaptive control methods can be grouped into three types, which are the gain scheduling, self tuning, and model reference. The gain scheduling method generally utilizes a look-up table to determine the best control parameters according to some known finite conditions of the plant. This method is easy to implement and fast but required a detailed knowledge of the plant to obtain the look-up table. The self tuning adaptive control continuously updates the model of the actual controlled system to be able to calculate the optimal parameter of the controller. It needs an elaborate model of the actual plant and complex calculation to obtain the optimal control parameters. The model reference compares the desired system response to the actual system response to adjust the controller's parameters. This method allows the non-adaptive control to work independently in case the adaptive algorithm fails.

The proposed adaptive proportional-integral-derivative (PID) controller is based on the model reference method, where the SPSA algorithm is used to adapt the PID gains. The detailed configuration of the proposed controller is shown in figure 1.

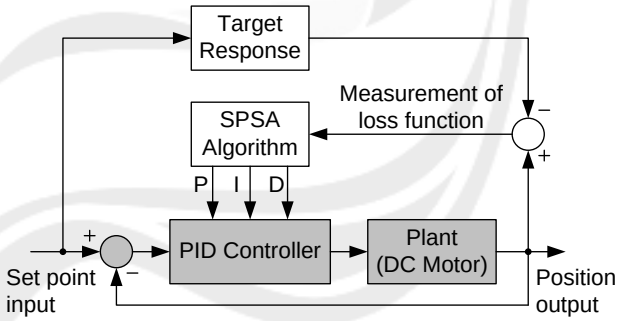


Figure 1. Block diagram of the proposed adaptive PID controller using SPSA algorithm.

The shaded components in the diagram are a general construct of a traditional PID controlled system. They are implemented in hardware with a direct current (dc) motor to represent the plant and a 16-bit PID controller. The 16-bit controller is implemented in a field programmable gate array (FPGA) [4]. The other components in the block diagram are implemented in a computer (PC) to add adaptive capability to the traditional PID control system.

The target response is a predefined desired position response of the dc motor modeled as a linear time invariant equation, i.e.,

$$\phi(t) = \phi_0 \exp\left(\frac{-t}{\tau}\right) + \phi_S\left(1 - \exp\left(\frac{-t}{\tau}\right)\right). \quad (1)$$

where ϕ_0 is the initial position of the dc motor and ϕ_S is the set point input of the system. The target response expressed in (1) is considered more realistic because the position of the dc motor changes gradually unlike a step function used in [3].

The loss function is measured as the integral of the absolute different between the target response and the actual position response over time, as shown as the shaded area in figure 2. As the shaded areas get smaller, the closer the actual response to the desired target response.

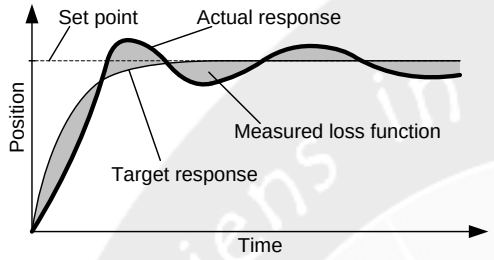


Figure 2. Loss function measurement example

3. THE 16-BIT PID CONTROLLER

Implementation of the proportional-integral-derivative (PID) controller in hardware is chosen for better performance in terms of sampling rate, and latency. It also shows that the adaptive method can be seen as a complementary add-on module to upgrade the existing traditional PID controller. The choice of 16-bit fixed point arithmetic is important to have an adequately precise controller.

The close loop position control system with PID is modeled as a continuous transfer function *F(s)*:

$$F(s) = \frac{M(s)}{E(s)} = \frac{D \cdot s^2 + P \cdot s + I}{s} \tag{2}$$

where *P*, *I*, and *D* are the proportional, integral, and derivative gain respectively, *M(s)* is the control output, and *E(s)* is the control input. The derived discrete transfer function *F(z)* of (2) is expressed as:

$$F(z) = \frac{M(z)}{E(z)} = \frac{b_0 + b_1 \cdot z^{-1} + b_2 \cdot z^{-2}}{1 - z^{-1}} \tag{3}$$

where:

$$\begin{aligned} b_0 &= P + \frac{I \cdot T_s}{2} + \frac{D}{T_s} \\ b_1 &= -P + \frac{I \cdot T_s}{2} - \frac{2 \cdot D}{T_s} \\ b_3 &= \frac{D}{T_s} \end{aligned} \tag{4}$$

and T_s is the sampling period.

The discrete time function can be directly derived from (3) as:

$$M(t) = b_0 \cdot E(t) + b_1 \cdot E(t - T_s) + b_2 \cdot E(t - 2T_s) + M(t - T_s) \tag{5}$$

where *M(t)* is the PID control output to drive the dc motor, *E(t)* is the different between the set point and the angular position of the dc motor. Equation (5) is implemented in hardware as shown in the following figure. The coefficient b_0, b_1, and b_2 are precalculated in the PC based on the perturbated PID gains.

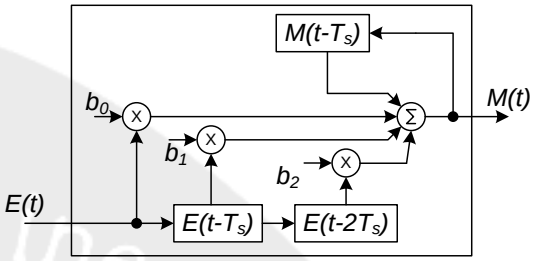


Figure 3. Hardware implementation of the PID controller

The multiplication, addition, and storage in figure 3 are all implemented as 16-bit fixed point digital circuits. The *M(t)* is however divided by 256 before being sent into an 8-bit pulse width modulation (PWM) module. High precision PWM module and dc motor is not necessary in this evaluation. Furthermore, the minimum magnitude of the output duty cycle of the PWM module is set to 58% to overcome friction in the mechanic of the dc motor.

Shaft position encoder and decoder logic provide a 16-bit angular position feedback from the dc motor.

Serial communication circuit included in the design provides interface between the PID controller and the SPSA algorithm in the PC. The serial baud rate is set at 38400 bauds per second. Values such as the position feedback, control output and the internal registers of the PID controller are sent to PC through the serial communication path. On the other direction, the PC can send the values set point, PID gains, and commands to control the PID controller.

4. THE SPSA ALGORITHM

The simultaneous perturbation stochastic approximation (SPSA) is a relatively new multivariate optimization algorithm [1]. This stochastic approximation algorithm is suitable for optimizing objective function that has many parameters, and where the objective function gradient is difficult or impossible to obtain. Other interesting features of the SPSA algorithm are the relatively small number of measurement required per iteration regardless of the dimension of the problem.

The SPSA algorithm does not need the gradient information of the objective functions, instead it approximates the gradient based on two perturbations and measurements *y(·)* of the loss function *L(·)*. These two measurements are made by simultaneously varying all of the parameters $\hat{\theta}$ in the problem. The measurement of the loss function generally includes additive noise, i.e. $y(\hat{\theta}) = (L(\hat{\theta}) + noise)$, where in this application the parameters are the PID gains:

$$\hat{\theta} = [\,\mathrm{P}\;\;\mathrm{I}\;\;\mathrm{D}\,]^{\mathrm{T}} \tag{6}$$

There are two types of gradient approximation methods. The first is the one-sided gradient approximation where the gradient is

approximated from two measurements, which are the measurement with the current set of parameter $y(\hat{\theta})$ and the measurement with the perturbated set of parameters $y(\hat{\theta} + perturbation)$. The second is the two-sided gradient approximation where the gradient is approximated from measurement with the positive and negative perturbation of the current set of parameters, i.e. $y(\hat{\theta} + perturbation)$ and $y(\hat{\theta} - perturbation)$ respectively. This paper uses the later method, where the positive perturbation $y(\hat{\theta}_k + c_k\hat{\Delta}_k)$ is called the "yplus" while the perturbated parameter set $(\hat{\theta}_k + c_k\hat{\Delta}_k)$ is called the "θplus", and the negative perturbation $y(\hat{\theta}_k - c_k\hat{\Delta})$ is called the "yminus" while the perturbated parameter set $(\hat{\theta}_k - c_k\hat{\Delta}_k)$ is called the "θminus", where k is the iteration number starting from 1 to the number of iteration.

The gradient for each parameter is approximated according to:

$$g_{kP}(\hat{\theta}_k) = \frac{y(\hat{\theta}_k + c_k\hat{\Delta}_k) - y(\hat{\theta}_k - c_k\hat{\Delta}_k)}{2c_k\Delta_{kP}} \quad (7)$$

$$g_{kI}(\hat{\theta}_k) = \frac{y(\hat{\theta}_k + c_k\hat{\Delta}_k) - y(\hat{\theta}_k - c_k\hat{\Delta}_k)}{2c_k\Delta_{kI}} \quad (8)$$

$$g_{kD}(\hat{\theta}_k) = \frac{y(\hat{\theta}_k + c_k\hat{\Delta}_k) - y(\hat{\theta}_k - c_k\hat{\Delta}_k)}{2c_k\Delta_{kD}} \quad (9)$$

or simplified to:

$$\hat{g}_k(\hat{\theta}_k) = \begin{bmatrix} g_{kP}(\hat{\theta}_k) \\ g_{kI}(\hat{\theta}_k) \\ g_{kD}(\hat{\theta}_k) \end{bmatrix} \quad (10)$$

The gain sequence c_k is a positive number that gets smaller as k gets larger according to:

$$c_k = \frac{c}{k^{\gamma}} \quad (11)$$

where the coefficient c sets the maximum gain of the perturbation, and the coefficient γ defines the exponentially decaying rate of the gain sequence c_k. The delta $\hat{\Delta}$ is an independently generated random vector with Bernoulli ±1 distribution, where each possible outcome has a probability of 0.5.

$$\hat{\Delta}_k = \begin{bmatrix} \Delta_{kP} \\ \Delta_{kI} \\ \Delta_{kD} \end{bmatrix} \quad (12)$$

The next set of PID gains is updated at the end of iteration k with the following relationship:

$$P_{k+1} = P_k - a_{kP}g_{kP}(\hat{\theta}_k) \quad (13)$$

$$I_{k+1} = I_k - a_{kI}g_{kI}(\hat{\theta}_k) \quad (14)$$

$$D_{k+1} = D_k - a_{kD}g_{kD}(\hat{\theta}_k) \quad (15)$$

where the gain sequences a_{kP}, a_{kI}, a_{kD} are decaying as k gets larger according to the equation:

$$\hat{a}_k = \frac{\hat{a}}{(k+A)^{\alpha}} \quad (16)$$

where:

$$\hat{a} = [\, a_P \;\; a_I \;\; a_D \,]^{\mathrm{T}} \quad (17)$$

The coefficients $\hat{a}$ set the maximum gain for the update of P, I, and D parameters individually, while the coefficient α determine the exponentially decaying rate of the gain sequences $\hat{a}_k$. The coefficient A increases the stability by neglecting the first A set of gain sequences $\hat{a}_k$ to compensate for a more aggressive values of $\hat{a}$.

The choice of gain sequences (c_k, and $\hat{a}_k$), along with the coefficients $\hat{a}$, c, and A govern the performance and stability of the SPSA algorithm, as with all stochastic approximation algorithms with their respective coefficients.

5. EVALUATION

For consistent evaluation result, the discrete time index k and the actual angular position of the dc motor is always reset to zero, and the PID controller is always reset to its initial state at the start of every perturbation. The set point ϕ_S is fixed at 0.5 all the time. At each iteration the output position response over time is recorded twice during the yplus and yminus perturbations, and then the PID gains $\hat{\theta}$ are updated based on the approximated gradients $\hat{g}_k(\hat{\theta}_k)$.

The gain sequences and coefficients of the SPSA algorithm are determined partly by trial and error with the guidance from [2] to find the suitable values for this implementation. Their complete values are shown in the following table.

Table 1. SPSA algorithm gain sequences, coefficients, and evaluation parameters

SPSA coefficients c	0.004
SPSA coefficient γ	0.5
SPSA coefficients [a_P, a_I, a_D]	[0.02, 0.01, 0.0003]
SPSA coefficient A	2
SPSA coefficient α	0.7
Initial proportional gain P_0	0.5
Initial integral gain I_0	0
Initial derivative gain D_0	0
Number of sample	3000
Number of iteration	20
Sampling rate T_s	45 Hz

Figure 4 shows the final position response after 20 iterations of SPSA algorithm, compared to the desired target response and the initial response. The initial position response in figure 4 is the output position response of the PID controller with the initial PID gains in table 1 above. It can be seen that the adaptive PID

controller managed to reduce the steady state error and reduce the rise time close to the desired target response without causing severe overshoot or oscillation to the final response.

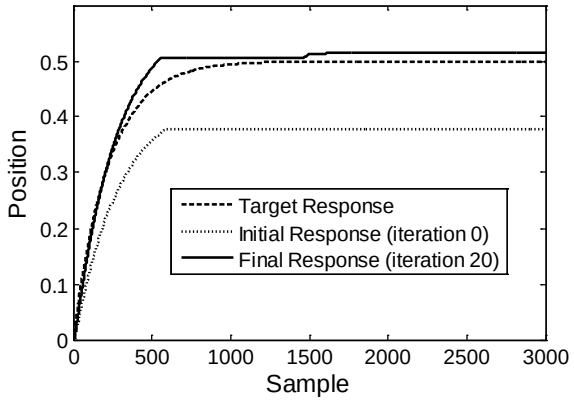


Figure 4. Position response of the adaptive PID controller before and after adaptation

Figure 5 shows the measured loss function for both the yplus and yminus perturbation at each iteration. The measured loss functions in figure 5 are normalized to the number of sample. The loss functions have approximately reached a saturated condition starting at the tenth iteration where both the loss functions are well below 0.025.

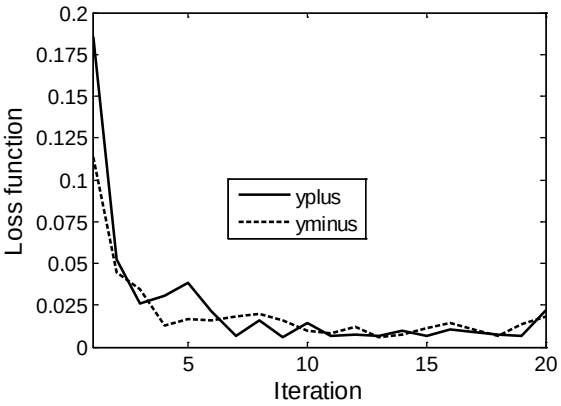


Figure 5. Yplus and yminus loss function measurement

The results in figure 5 are supported in figure 6, where all the PID gains start to converge at the tenth iteration.

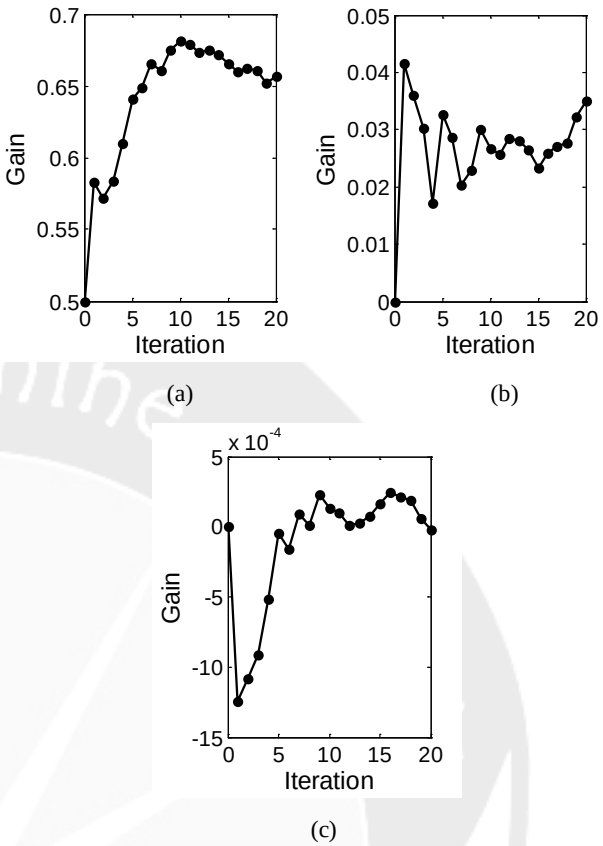


Figure 6. Values of (a) the proportional gain, (b) the integral gain, and (c) the derivative gain

Furthermore by having individual SPSA coefficient a for each PID gain in equation (13-15), each PID gain update can have different step size per iteration. This is suitable for the simple PID controller that is more sensitive to the change in proportional and integral gains compared to the derivative gain.

6. CONCLUSIONS

The proposed adaptive PID controller using the SPSA algorithm and realistic target response function has been shown to be able to adapt its own PID gains to minimize steady state error and reduce rise time, close to the desired target position response over time. The PID gains approximately start to converge at the tenth iteration.

Using the linear time invariant system as a model to the target response, the rise time of the output position response can be reduced without causing severe overshoot or oscillation.

The independent gain sequence a_k for each PID gains has been shown to update each PID gains with different step size per iteration.

The existing PID controller can be upgraded to an adaptive PID controller simply by tapping the SPSA algorithm to the input set point and output position, provided that the PID gains of the existing PID controller can be updated in real time.

7. REFERENCES

[1] Spall, James C. 1998. An overview of the simultaneous perturbation method for efficient optimization. Johns Hopkins APL Technical Digest, Vol 19 No 4.

[2] Spall, James C. 1998. Implementation of the simultaneous perturbation algorithm for stochastic optimization. IEEE Transactions on Aerospace and Electronic Systems, Vol 34 No 3.

[3] Tan, Sofyan and Hian, Lie. 2009. Kontrol motor PID dengan koefisien adaptif menggunakan algoritma simultaneous perturbation. Proceeding Konferensi Nasional Sistem dan Informatika (STMIK Stikom Bali, Indonesia, November 14, 2009)

[4] Tan, Sofyan. 2005. Sistem kontrol PID 16-bit menggunakan FPGA. Seminar Nasional: Soft Computing, Intelligent Systems and Information Technology (Universitas Kristen Petra, Surabaya, Indonesia, June 02, 2005)

Induction Heating Efficiency Analysis Modeling Using COMSOL® Multiphysics Software

Didi Istardi
Batam Polytechnics Parkway st, Batam Centre
Batam, Indonesia
+62-778-469856
istardi@polibatam.ac.id

ABSTRACT
Induction heating is clean environmental heating process due to a non-contact heating process. A lot of researches work in development of a new material and design heating process. With COMSOL Multiphysics software, the phenomena in induction heating process can be simulated and estimated. Therefore, the effect of inductor's width, inductor's distance, and conductive plate material in induction heating process were also simulated. The result shown that the efficiency of induction heating influenced by the width's variations and conductive material.

Keywords
COMSOL, efficiency, induction heating.

1. INTRODUCTION
A green and renewable energy, high cost, and clean environmental are important issues that influenced the technology in home appliances recently such as in stove and water heating. Induction heating is one of the new technologies in home appliance. There are some researches that discussed about induction heating process and implementation due to their clean and no pollution [1-5]. According to this reasons, the induction heating efficiency is simulated and analyzed using COMSOL Multiphysics software. In this paper, the effect of frequency, distance of inductor, and conductive material of inductor are also considered.

Induction heating is a non-contact heating process. It uses high frequency electricity to heat materials that are electrically conductive [6]. Since it is non-contact, the heating process does not contaminate the material being heated. It is also very efficient since the heat is actually generated inside the work piece. This can be contrasted with other heating methods where heat is generated in a flame or heating element, which is then applied to the work piece.

The paper is organized as follows: In the next section, a brief review of induction heating theory is presented. In Section III, a description of the problem setting in COMSOL is explained. Section IV explains the geometry object and constraint in this simulation. Section V presents results of simulation using COMSOL Multiphysiscs software and discussion of the results. Finally, the conclusions are made in section VI.

2. INDUCTION HEATING BASIC
A source of high frequency electricity is used to drive a large alternating current through a coil. This coil is known as the work coil. The passage of current through this coil generates a very intense and rapidly changing magnetic field in the space within the work coil. The work piece to be heated is placed within this intense alternating magnetic field [6, 7].

The alternating magnetic field induces a current flow in the conductive work piece. The arrangement of the work coil and the work piece can be thought of as an electrical transformer. The work coil is like the primary where electrical energy is fed in, and the work piece is like a single turn secondary that is short-circuited. This causes tremendous currents to flow through the work piece. These are known as eddy currents.

In addition to this, the high frequency used in induction heating applications gives rise to a phenomenon called skin effect [6, 7]. This skin effect forces the alternating current to flow in a thin layer towards the surface of the work piece. The skin effect increases the effective resistance of the metal to the passage of the large current. Therefore it greatly increases the heating effect caused by the current induced in the work piece. The principle of induction heating is mainly based on two well-known physical phenomena:

Electromagnetic induction, the energy transfer to the object to be heated occurs by means of electromagnetic induction. It is known that in a loop of conductive material an alternating current is induced, when this loop is placed in an alternating magnetic field. The formula is the following [7]:

$$E = \frac{d\phi}{dt} \tag{2.1}$$

When the loop is short-circuited, the induced voltage E will cause a current to flow that opposes its cause – the alternating magnetic field. This is Faraday - Lenz's law.

If a 'massive' conductor (e.g. a cylinder) is placed in the alternating magnetic field instead of the short circuited loop, than eddy currents (Foucault currents) will be induced in here. The eddy currents heat up the conductor according to the Joule effect

The second phenomena is joule effect, when a current I [A] flows through a conductor with resistance R [Ω], the power P [W] is dissipated in the conductor according to the formula [7].

$$P = RI^2 \tag{2.2}$$

A general characteristic of alternating currents is that they are concentrated on the outside of a conductor. This is called the skin effect. Also the eddy currents, induced in the material to be heated, are the biggest on the outside and diminish towards the centre. So, on the outside most of the heat is generated. The skin effect is characterized by its so-called penetration depth d. The penetration depth is defined as the thickness of the layer, measured from the outside, in which 87% of the power is developed. The penetration depth can be deduced from Maxwell's equations. For a cylindrical load with a diameter that is much bigger than δ , the formula is as follows:

$$\delta = \sqrt{\frac{\rho}{\pi.\mu.f}} \tag{2.3}$$

The penetration depth, on one hand, depends on the characteristics of the material to be heated (μ, ρ) and, on the other hand, is also influenced by the frequency. The frequency dependence offers a possibility to control the penetration depth. As can be derived from the formula above, the penetration depth is inversely proportional to the square root of μr. For non-magnetic materials like copper or aluminum the relative magnetic permeability is μr=1. Ferromagnetic materials (iron, many types of steel) on the contrary, have a μr-value that is much higher. Therefore, these materials generally show a more explicit skin effect (smallerδ). The magnetic permeability of ferromagnetic materials strongly depends on the composition of the materials and on the circumstances (temperature, magnetic field intensity, saturation). Above the Curie temperature μr suddenly drops again to μr=1, which implies a rapid increase of the penetration depth. The current flow in skin effect can be calculated using equation:

$$i_x = i_o e^{-x/\delta} \tag{2.4}$$

Where, i_x is distance from the skin (surface) of the object, current density at x and Io refer to current density on skin depth.

3. PROBLEM SETTING

In this paper, the quantities that want to be computed and analyzed are the effect of the inductor's width and distance on the efficiency at frequency 50 Hz and 2 kHz and the material of the plate affects the results. According to this problem, the equations to be solved in the COMSOL Multiphysiscs software will be start with Ampere's law as seen in,

$$\nabla xH = J + \frac{\partial D}{\partial t} = \sigma E + \sigma vxB + J^c + \frac{\partial D}{\partial t} \tag{3.1}$$

Now assume time-harmonic fields and use the definitions of the potentials,

$$B = \nabla xA \tag{3.2}$$

$$E = -\nabla V - \frac{\partial A}{\partial t} \tag{3.3}$$

Combine them with the constitutive relationships B = μ_0 (H + M) and D = ε_0 E + P to rewrite Ampere's law as

$$\left(j\omega\sigma - \omega^2\varepsilon_o\right)A + \nabla x\left(\mu_o^{-1}\nabla xA - M\right) - \sigma vx(\nabla xA) + \left(\sigma + j\omega\varepsilon_o\right)\nabla V = J^c + j\omega P \tag{3.4}$$

In the 2D in-plane case there are no variations in the z direction, and the electric field is parallel to the z-axis. Therefore you can write ∇V as −ΔV/L where ΔV is the potential difference over the distance L. Now simplify these equations to

$$-\nabla.\left(\mu_o^{-1}\nabla A_z - \begin{bmatrix} -M_y \\ M_x \end{bmatrix}\right) + \sigma v.\nabla A_z + \left(j\omega\sigma - \omega^2\varepsilon_o\right)A_z = \sigma\frac{\Delta V}{L} + J_z^2 + j\omega P_z \tag{3.5}$$

The mathematical model for heat transfer by conduction is the following version of the heat equation

$$\delta_{ts}\rho C_p\frac{\partial T}{\partial t} - \nabla.(kVT) = Q \tag{3.6}$$

The resistance per meter is defined as $R = \frac{P}{I^2}$ where P is the power loss, and I is the current through the inductor and can be calculated using the formula:

$$P = \int Q.dA \tag{3.7}$$

$$I = \int I_z dA \tag{3.8}$$

The value of the second integral is known, because the total current is part of the boundary conditions. However, performing the integration of the FEM solution provides an opportunity to verify that the previous calculations are correct.

The skin depth, that is, the distance where the electromagnetic field has decreased by a factor e−1, is for a good conductor,

$$\delta = \sqrt{\frac{2}{\omega\mu\sigma}} = \sqrt{\frac{\rho}{\pi\mu f}} \tag{3.9}$$

4. OBJECT GEOMETRY AND MATERIAL

The geometry of this simulation in this paper can be seen in Figure 1.

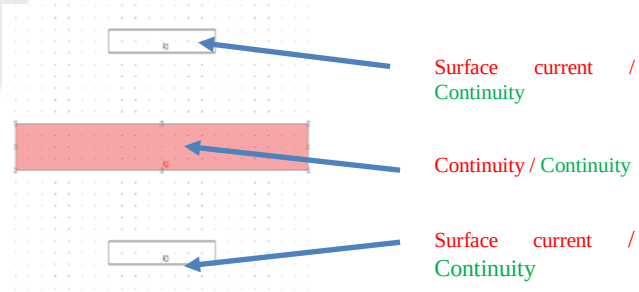


Figure 1. Object geometry of induction heating

Information written in red refer to the AC Power Electromagnetic (q_a) mode, and the ones in green refer to the Heat Transfer by

Conduction (h_t) mode. The heat source comes from the resistive heating, time average (Q_{av_qa}) which is used inside the conductive plate.
Material of induction heating referred to Table 1.

Table 1. Material properties of induction heating

Material	Electric Conductivity (σ)	Permeability (μ)	Heat capacity (J/kg.K)	Therm. conductivity	Density (kg/m^3)
Aluminum	$3.774.10^7$	1	900	160	2700
Copper	$5.998.10^7$	1	385	400	8700
Iron	$1,12.10^7$	4000	400	76,2	7870
Steel	$4,032.10^6$	1	475	44,5	7850
Silicon Carbide	1.10^3	1	1200	$450(300/T)^{0..75}$	3200

Sub domain settings of this simulation can be referred to Table 2 and Table3.

Table 2. Properties of AC Power Electromagnetic (q_a)

Settings		Sub domain 1, 3-4	Sub domain
Relative permeability	1	{1,0;0,1}	{1,0;0,1} (Copper)
Relative permittivity	1	1	1 (Copper)
Electric conductivity	S/m	0	5.998e7[S/m] (Copper)

Table 3. Properties of Heat Transfer by Conduction (h_t)

Settings	Sub domain 1	Sub domain 2	Sub domain 3,4
Thermal conductivity	0.026	400[W/(m*K)]	400[W/(m*K)]
Density	1.293	8700[kg/m^3]	8700[kg/m^3]
Heat capacity	$1.01e^3$	385[J/(kg*K)]	385[J/(kg*K)]
Heat source	0	Q_{av_qa}	0

5. RESULTS AND DISCUSSION

5.1 Inductor's Width

In this part of the report, the inductor's width has been changed, for two different frequencies 50 Hz and 2 kHz. First of all, the skin effect which has been introduced before should be presented using a practical case and the software COMSOL. The following example compares the temperature and current density distributions for the same configuration; only the frequency is changed (50 Hz and 2 kHz) as can be seen in Figure 2.

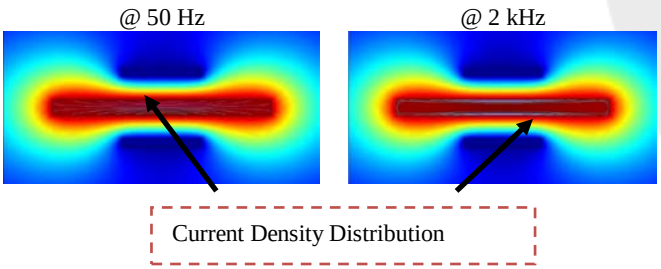


Figure 2. Current density distribution in different frequency

It is easy to see how the skin effect affects the current distribution. With a higher frequency, the current density is increased near the boundary, and the different penetration depth (δ) can be calculated in order to quantify this skin effect.

From the Figure 2, the skin effect does not effect to the variation of the inductor's width. Skin effect only depends on frequency, electric conductivity and permeability (equation 2.3). Based on this equation, the values of skin effect can be seen in Table 4.

Table 4. Comparison of skin depth

Skin depth(mm)	Frequency	
	50 Hz	2 kHz
Calculate	3	0.5
Simulation	8	2

In the following analyze, the width's of the inductor has been increased, but the distance between the inductor winding and the conductive plate has been kept constant. The result of simulation can be seen in Figure 3.

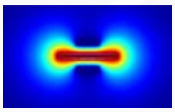
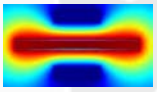
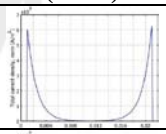
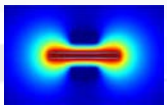
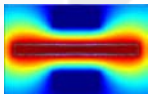
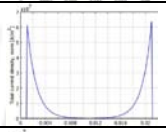
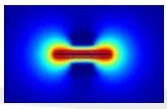
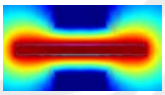
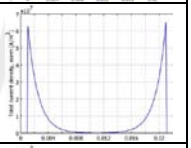
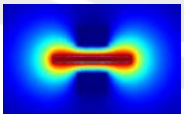
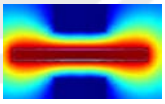
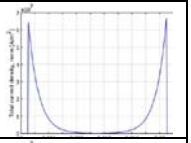
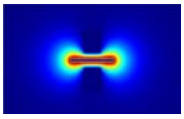
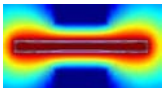
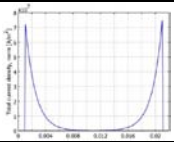
Width (m)	Temperature Distribution (2 kHz)	Current density distribution (2 kHz)	Total Current Density distribution (2 kHz)
0,02			
0,03			
0,04			
0,05			
0,1			

Figure 3. Effect of Inductor's width

At frequency 2kHz, the current distribution is not linear anymore [cf. Equation 2.4]. The current density is higher near the boundary (skin effect), and it can be seen that the increasing of width amplify this phenomena.

The density of the magnetic field wanes as the object gets closer to the center from the surface. According to Faraday's Law, the current generated on the surface of a conductive object has an inverse relationship with the current on the inducting circuit as

described in [cf. Equation 2.4].The current on the surface of the object generates an eddy current .

If an object has conductive properties like iron, additional heat energy is generated due to magnetic hysteresis. The amount of heat energy created by hysteresis is in proportion to the size of the hysteresis. The efficiency can be calculated using equation:

$$\mu = \frac{P_{out}}{P_{in}} = \frac{P_{out}}{i^2 R_{copper}} \quad (5.1)$$

P_{out} can find from postprocessing and Subdomain integration (integration the power in average time) with subdomain in plate material and in P_{in} in subdomain inductor.

If the inductor's width changes the current density also will change

$$J_{s0} = \frac{I_0}{A} \quad (5.2)$$

where the A are area of inductor. So, if area of inductor increases, the current density will be decreased. Therefore, the electromagnetic induces to plate material also reduce and the power in plate material also reduced (Ampere's Law), and the efficiency will decrease.

If the frequency increase the power will be increase also the efficiency is higher than in lower frequency. It's can be illustrate that in high frequency, the ac resistance (R_{ac}) will be higher than dc resistance (R_{dc}), it's happen due to skin effect phenomena.

$$\frac{loss}{surface} = \left(\frac{I_{RMS}}{width}\right)^2 R <=> R = \frac{\rho}{\delta} \quad (5.3)$$

$$\eta = \frac{R}{R + R_{in}} \quad (5.4)$$

The comparison of efficiency and power output in different width and frequency can be seen in Table 5.

Table 5. Efficiency and power output of induction heating in different frequency

Width (m)	Efficiency		Power Output (kW)	
	50 Hz	2 kHz	50 Hz	2 kHz
0.02	41.74%	97.16%	4594	10694
0.03	40.82%	96.16%	2252	5230
0.04	38.13%	88.53%	1387	3221
0.05	34.54%	80.19%	965	2240
0.1	14.62%	34.07%	335	782

5.2 Inductor's Distance

The distance between inductor and conductive plate also affect the induction heating process, and this mechanism is called proximity effect. If the distance between inductor plate and conductive plate increase, then the strength of the proximity effect will decrease. Due to Faraday's Law, the eddy current within the conductive plate have an opposite direction to that of the source current of inductor. Therefore, due to proximity effect, the inductor current and the conductive plate eddy current will concentrate in the area facing each other.

For example, at 50 Hz, it can be done by increasing the distance between the inductor and the plate, but the width of the inductor stayed constant as shown in Figure 4.

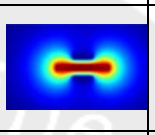
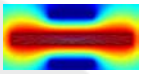
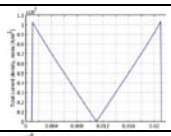
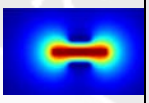
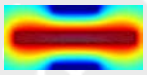
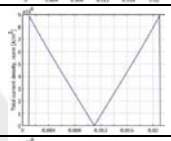
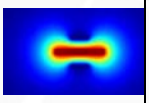
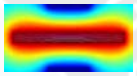
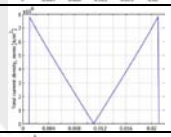
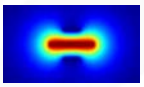
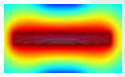
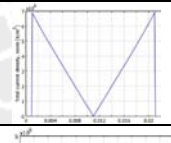
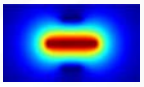
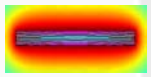
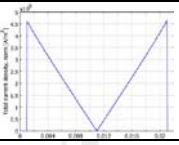
Distance	Temperature Distribution (50 Hz)	Current density distribution (50 Hz)	Total Current Density distribution (50Hz)
0,03 m			
0,04 m			
0,05 m			
0,06 m			
0,1 m			

Figure 4. Effect of inductor's distance

It can be observed that to change the distance between the inductor and the plate, has consequences on the temperature distribution outside the material but not on the skin effect. Actually, the current density is still linear inside the material. However, it can also be noticed that the efficiency has changed.

The eddy current has a higher density in the conductive plate area where the distance between them (air gaps) is small. It means, they have a good coupling. Therefore, there will be intense heating due to the Joule effect.

Table 6. Maximum current density in conductive plate

Distance of air gaps	Maximum Current density [A/m²]
0.03 m	1.02.107
0.04 m	9.106
0.05 m	8.106
0.06 m	7.106
0.1 m	5.106

It can be seen on this part that the current distribution is not disturbed by the distance between the inductor and the conductive

plate, but the efficiency of the induction heating is affected. It is easy to understand that the heating action is decreased when the distance increases.

5.3 Conductive Plate Material

At last but not least, an interesting study is about the variations of the material of the plate. The following example shows a comparison between copper, iron, steel and silicon carbide as seen in Figure 5.

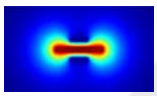
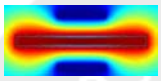
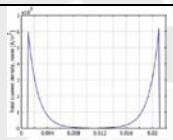
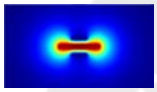
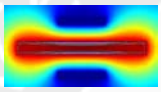
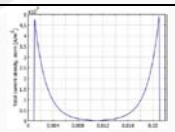
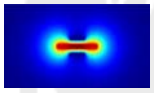
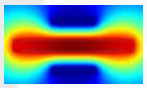
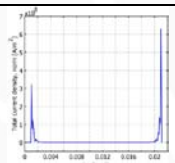
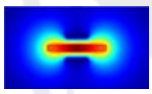
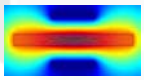
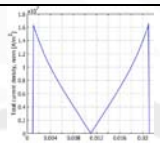
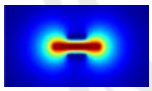
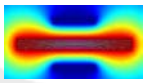
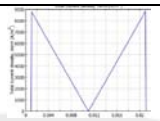
Material	Temperature Distribution (50 Hz)	Current density distribution (50 Hz)	Total Current Density distribution (50Hz)
Copper			
Al			
Fe			
Steel			
Sic			

Figure 5. Effect of conductive plate material

It can be shown that the distribution of the current density inside the material is more or less linear, depending on the material used. The current distribution inside the copper or the aluminum are strongly non linear, but it is almost linear if the material used is steel, or even more with silicon carbide. This is related to the electric conductivity. The different values of electric conductivity for each material can be compared as follow: $\sigma_{cu} > \sigma_{Al} > \sigma_{Steel} > \sigma_{Sic}$; so, smaller the value of the electric conductivity is, more linear the current density inside the material will be.

However, there is an exception to this behavior, it is iron. Actually, the material has an unexpected behavior, but this is due to its relative permeability which is much higher than for the other materials. The comparison of different material can be seen in Table 7.

Table 7. Effect of conductive plate material in induction heating

Material	Skin depth (mm)		Temperature after 1020 sec, in °K
	Calculate	Simulation	
Aluminum	0,58	2,5	515,5
Copper	0, 5	2	590,5
Iron	0,0168	0,5	14 816
Steel	1,177	8	1464,5
Silicon	112	10	294,5

Table 5 shows that the penetration depth and the temperature will be varied with the electric conductivity and permeability.

6. CONCLUSION

It has been shown that the skin effect (non linearity of the current distribution) is affected by the frequency, and the material used, but it is affected neither by the conductor width nor by the distance between conductor and conductive plate.

The width variations have no effect on the current density distribution, but it affects the efficiency. One was that if the width is increased, the efficiency is then reduced. For the distance variations, the current density distribution is not effected in this case also, only the maximum value who is decreased when the distance is increased.

If the conductive plate material is changed, variations on the current density distribution have been observed. Actually, if we increase the electric conductivity, the linearity of the distribution is decreased.

7. ACKNOWLEDGMENTS

Thanks to Nicolas Rouch for the discussion and suggestion and also to Chalmers University of Technology, SWEDEN that allow me to use the COMSOL Software in this project.

8. REFERENCES

[1] Ernest, R., Perrier, D., Feigenblum, J., and Hemous, R. 2006. 3D Inductive Phenomena Modelling. In The Proceeding of the COMSOL Users conference 2006 Paris.

[2] Savia, D. J. 2007.Induction Heating of Samples in Vacuum Systems. In The Proceeding of the COMSOL Users conference 2007 Grenoble.

[3] Jungwirth, M. and Hofinger, D. 2007. Multiphysics Modelling of High-Frequency Inductive Device. The Proceeding of the COMSOL Users conference 2007 Grenoble.

[4] Jankowski, T. A. and Others. 2009. Experimental Observation and Numerical Prediction of Induction Heating in Graphite Test Article. In The Proceeding of the COMSOL conference 2009 Boston.

[5] Wislati, R. and Haase, H. 2007. Using COMSOL Multiphysics for the Modelling of a hybrid Linear Stepper

Motor. In The Proceeding of the COMSOL Users conference 2007 Grenoble.

[6] Rudnev V. I. 2003. Handbook of Induction Heating. Marcel Dekker Inc.

[7] Leinhard IV, J H and Leinhard V, J.H. 2006. A Heat Transfer Textbook.

Authors Index

ICSIIT

2nd International Conference on Soft Computing, Intelligent System and Information Technology 2010

ISBN: 978 602-97124-0-7

Organized by:

Informatics Department
Petra Christian University
Siwalankerto 121-131 Surabaya
http://icsiit.petra.ac.id
http://informatika.petra.ac.id