Volume 10 Issue 2 June 2017 ISSN 2088-7051

Jurnal Ilmu Komputer dan Informasi

Journal of Computer Science and Information

Jurnal Ilmu Komputer dan Informasi (Journal of a Science and Information). 10/2 (2017), 59-66
DOI: http://dx.doi.org/10.21609/jiki.v10i2.374

AUTOMATIC ONTOLOGY CONSTRUCTION USING TEXT CORPORA AND ONTOLOGY DESIGN PATTERNS (ODPS) IN ALZHEIMER'S DISEASE

Denis E. Cahyani[1] and Ito Wasito[2]

[1]Departement of Informatics, Faculty of Mathematics and Natural Sciences, Universitas Sebelas Maret, Jl. Ir. Sutami No.36A, Jebres, Kota Surakarta, Jawa Tengah 57126, Indonesia.
[2]Faculty of Computer Science, Universitas Indonesia, Kampus UI, Depok, 16424, Indonesia

E-mail: denis.eka@staff.uns.ac.id, ito.wasito@cs.ui.ac.id

Abstract

An ontology is defined as an explicit specification of a conceptualization, which is an important tool for modeling, sharing and reuse of domain knowledge. However, ontology construction by hand is a complex and a time consuming task. This research presents a fully automatic method to build bilingual domain ontology from text corpora and ontology design patterns (ODPs) in Alzheimer's disease. This method combines two approaches: ontology learning from texts and matching with ODPs. It consists of six steps: (i) Term & relation extraction (ii) Matching with Alzheimer glossary (iii) Matching with ontology design patterns (iv) Score computation similarity term & relation with ODPs (v) Ontology building (vi) Ontology evaluation. The result of ontology composed of 381 terms and 184 relations with 200 new terms and 42 new relations were added. Fully automatic ontology construction has higher complexity, shorter time and reduces role of the expert knowledge to evaluate ontology than manual ontology construction. This proposed method is sufficiently flexible to be applied to other domains.

Keywords: *fully automatic, ontology building, ontology design patterns, Alzheimer disease*

Abstrak

Ontologi didefinisikan sebagai spesifikasi eksplisit dari sebuah konseptualisasi, yang merupakan alat penting untuk pemodelan, pembagian, dan penggunaan kembali pengetahuan domain. Namun, konstruksi ontologi dengan tangan merupakan tugas yang rumit dan memakan waktu. Penelitian ini menyajikan metode otomatis untuk membangun ontologi domain bilingual dari pola desain korporat teks dan ontologi (ODPs) pada penyakit Alzheimer. Metode ini menggabungkan dua pendekatan: pembelajaran ontologi dari teks dan sesuai dengan ODP. Ini terdiri dari enam langkah: (i) ekstraksi istilah & hubungan (ii) Pencocokan dengan glosarium alzheimer (iii) Pencocokan dengan pola desain ontologi (iv) Perhitungan skor kesamaan istilah & hubungan dengan ODPs (v) Ontologi bangunan (vi) Evaluasi Ontologi. Hasil ontologi yang terdiri dari 381 istilah dan 184 hubungan dengan 200 istilah baru dan 42 hubungan baru ditambahkan. Konstruksi ontologi otomatis lengkap memiliki kompleksitas yang lebih tinggi, waktu yang lebih singkat dan mengurangi peran pengetahuan ahli untuk mengevaluasi ontologi daripada konstruksi ontologi manual. Metode yang diusulkan ini cukup fleksibel untuk diterapkan pada domain lain.

Kata Kunci: *fully automatic, ontology building, pola desain ontologi, penyakit alzheimer*

1. Introduction

Alzheimer's disease is one of the important issues in the field of public health. In France, there are about 860,000 people affected by Alzheimer's disease. Each year there are 220,000 new cases identified in the country. To prevent the increase in patients suffering from Alzheimer's disease and addressing existing cases, clinical epidemiology training required for health workers. The purpose of the training is to be able to improve the quality of health care workers and to increase the number of health workers who are able to contribute to handle cases that occur in Alzheimer's disease. The improvement of quality medical practice needed to prevent the increase of Alzheimer's disease's patient. One of way to prevent this is educating practitioners in clinical epidemiology. So the number of medicinal practice that can be handle cases about Alzheimer's disease increasing.

Conversely, rapid and efficient decision making is a crucial issue in the public health domain and especially in the Alzheimer's disease domain.

Decision-makers should refer to various experts opinions as they cannot screen themselves all the scientific facts reported in different sources including online scientific literatures and results of clinical trials.

Within this framework, there is a corpora named BiblioDem Digital Library. BiblioDem Digital Library is a critical review of scientific papers related to Alzheimer's disease from a variety of reference journals. Critical analysis of the various articles are reviewed by a domain expert research and integrated into an online bibliographic database called BiblioDem. There are 6-11 relevant papers selected for publication in the monthly magazine named BiblioDemences. BiblioDem contains over 1500 documents which contain title, abstract, critical analysis and the name of experts who carried out the analysis. The paper will be published in BiblioDem if the query contains a term which is shown as the title or abstract in the paper.

Beside using a BiblioDem text corpora, this research also uses ontology design patterns. Ontology design patterns (ODPs) is derivative of the design patterns which used in software engineering. Ontology design patterns is a pattern that makes it possible to identify the design of the ontology structure. Design patterns allow for the regulation of inter-term dependencies so if there is a change in the term it will not affect the other terms. Examples of types of ODPs is extensional patterns, good practice patterns, modeling patterns that can be implemented using the OWL format [1].

This research also uses the Alzheimer glossary to filter word extracted from text corpora BiblioDem. Glossary Alzheimer's is a list of vocabulary and their definitions related to Alzheimer's disease. A glossary contains explanations of concepts relevant to a particular topic and related to the ontology.

This research will be conducted in bilingual domain ontology construction using a text corpus and matching with ontology design patterns for representing knowledge through ontology. In this research, the ontology will be built automatically, which aims to reduce the role of human or expert knowledge to build ontology.

Related Work

An ontology is defined as an explicit specification of a conceptualization, which is an important tool for modeling, sharing and reuse of domain knowledge [2]. It allows domain knowledge to be represented explicitly through concepts and relations between them and hence to manipulate it automatically. However, ontology construction by hand is a complex and time consuming task [3]. Therefore, an automatic process is needed to help to facilitate the construction of ontology. Existing example approach to automated process is ontology design patterns (ODPs) [4].

The research of development of semi-automatic ontology using existing resources had been developed previously. Drame et al [5] builds a semi automatic-multilingual domain ontology is using UMLS Metathesaurus and parallel corpus. Validation of the ontology is constructed using Alzheimer's disease expert to ensure ontology constructed in accordance with the knowledge in Alzheimer's disease. However, this validation takes about a month to validate the ontology. It is take a lot of time. Therefore, this study developed using ODPs that validation can be done without the help of an expert. It aims to accelerate the development process ontology. The proposed method is an extension of conventional semi-automatic method.

Three studies related to the automatic ontology construction is research that conducted by Dahab et.al [6] who build automated construction ontology from natural language text. Then, Chen et.al [7] using recursive adaptive resonance Training (ART) network to construct a domain ontology-based TF-IDF. The study using web pages to build an ontology automatically. Navigli and Velardi [8] develop methodology for automatic ontology enrichment and document annotation.

Dahab et al build the domain ontology of natural text that using semantic pattern-based approach. This study analyzes the natural domain text to extract candidate relations and terms and mapping it into ontology. Meanwhile, Chen et.al using the Internet and web pages using HTML tags labels to choose the terms of web pages. Then calculated the TF-IDF to find weights of the used terms and then use the network ART (Adaptive Resonance Theory Network) to cluster terms. In study that conducted Navigli and Velardi, natural language definitions from available glossaries are processed and regular expressions are applied. The purpose is to identify general-purpose and domain-specific relations. The process in this research consist of pre-processing step (part-of-speech tagging and Named Entity Recognition), annotation of sentence segments with CIDOC properties and formalization of glosses. The evaluation methodology performance is extracting hypernymy and non-taxonomic relations. This study assessed the generality of the approach on a set of web pages from the domains of history and biography. The research in this paper is different from the three studies before because in this study

using a bilingual text corpora as a material to build ontology and using ontology approach design patterns (ODPs).

2. Methods

Resources

The BiblioDem Corpus

BiblioDem is a cumulative bibliographic database which currently contains 1556 scientific papers on Alzheimer's disease and related syndromes. This database contains abstracts of scientific papers selected from worldwide literature on Alzheimer's disease and their associated critical analysis, thus constituting rich knowledge. There are two kinds of corpus, namely corpus in English language and French language.

The corpus used in this research differs from previous research corpus [5]. Previous corpus contains scientific papers from 2004-2011, whereas I n this research using a corpus of scientific papers in 2013-2014 which amounts to 125 papers. The corpus can be obtained at the website address http://sites.isped.ubordeaux2.fr/bibliodem/bulletins.aspx.

Alzheimer Glossary

This research using Alzheimer glossary for filter extracted term from text corpora. The filter-ing term extraction is necessary because this process can produce the terms that have special relation with Alzheimer's disease, not term which is concerned with general health. Glossary Alzheimer can be obtained at the website address http://alzheimers.about.com/od/glossary/ and http://www.webmd.com/alzheimers/glossaryterms alzheimers. The total of terms which are related to Alzheimer disease in the Alzheimer glossary are 370 terms.

Ontology Design Patterns (ODPs)

Ontology Design Patterns (ODPs) can be accessed at http://www.gong.manchester.ac.uk/odp/html/index.html. This website also contains a catalog of ODPs. In this catalog, there are three types of ODPs namely (i) Domain Modelling ODPs, (ii) Good Practice ODPs (iii) Extension ODPs. The total number of ontology design patterns in the catalog number 16 ODPs. ODPs Domain Modeling aims to get the best model for a domain specific ontology. For example, Interactor_Role_Interaction and Sequence. Good Practice ODPs ontology aims to get better and stronger to maintain ontology models. For example, Normalization and Upper Level Ontology. On the other hand, ODPs Extension aims to overcome the limitations of existing ontology models to expand or increase coverage of the ontology. For example, Nary_DataType Relationship and Exception.

Tools

Text2Onto

Text2Onto is a framework of learning ontology which developed to support ontology construction from textual documents. Text2Onto has been used Cimiano and Volker [3]. The research used Text2Onto as a framework for ontology learning from textual resources based on Probabilistic Ontology Model (POM). There are three processes in Text2Onto: preprocessing, Execution of Algorithms and Combining results. During preprocessing, Text2Onto calls GATE application to tokenize document and tag Part of Speech sentences to creates indexes for the document and the result of this process is obtained as an annotation document. Execution of Algorithms is the process of Text2Onto executes the applied algorithms to extract terms and relations. One of the applied algorithm is TFIDF Concept Extraction. The last process is combining results, this process combines result of extracted terms and relations derived from processed documents. Text2Onto can be accessed at http://code.google.com/p/text2onto/downloads/list.

SimMetrics

SimMetrics is an open-source library available in Java which contains more than 20 similarity distance algorithms. For example, Jaro-Winkler, Levenstein distance, and Monge Elkan distance. SimMetrics used for string matching to identify the position of string or set of strings within a text. String matching algorithms helps to compare two different strings and look for similarity score between two text comparison. SimMetrics has been used by Chapman et al [9]. This research using simMetrics to calculate similarity between texts, where the information in this text will be integrated in a large repository (e.g. the Web). SimMetrics can be accessed at https://github.com/Simmetrics/simmetrics.

Ontology Generation

Ontology generation is a plugin in protégé to build ontology with generate terms of natural language text. Ontology generation was developed by Watcher and Schroeder, 2010 [10]. This tool supporting the creation and extension of OBO ontology by semi-automatically generating terms, definitions and parent–child relations from text in PubMed, the web and PDF repositories. This tool generates term by identifying significant noun phrases in text statistically and for the definitions and parent–child relations it employs pattern-

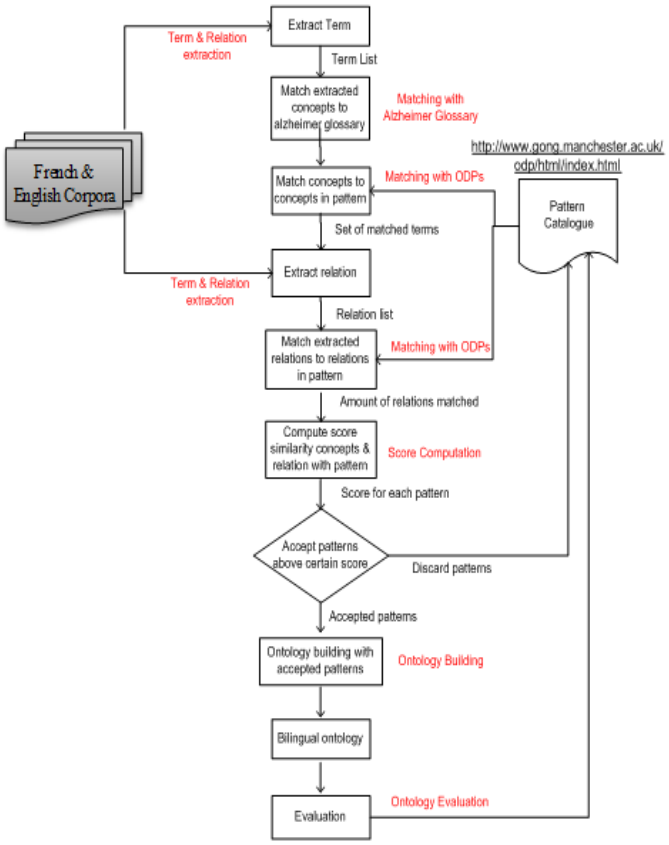


Figure 1. Overview of the ontology building method

based web searches. Ontology generation can be obtained at http://protegewiki.stanford.edu/wiki/Ontology_Generation_Plugin_(DOG4DAG). Ontology generation can be applied to the protégé-OWL version 4.1

Methods

The methods in this research consists of six stages: (i) Term and relation extraction (ii) Matching with Alzheimer glossary (iii) Matching the ontology design patterns (iv) Score computation similarity term and relations with ODPs (v) Ontology Building (vi) Ontology evaluation. To explain the process of each stage in the methodology is indicated in the figure 1.

The idea of the methodology is to take the extract-ed terms and relations, match them against the patterns and depending on the result use parts of the patterns to build the ontology. As a preprocessing step, a text corpus was analyzed by some term extraction software, which renders a list of possibly relevant terms. This list of terms is the input for this method.

Term & Relation Extraction

BiblioDem corpus extracted to obtain several terms. Further, relation which links between terms are extracted to obtain several relations. This corpus extraction using tools called Text2Onto.

Matching with Alzheimer Glossary

Having acquired several terms and relations, this terms and relation have matched with an Alzheimer glossary. At this stage, the matching with Alzheimer glossary is aimed to filter term so the same term derived from the extracted word list and list of words in the glossary.

Matching with ontology design patterns

The extracted terms and relations will be compared with terms and relations contained in Catalog ODPs that consist 16 design patterns. The matching result will be calculated score of similarity by SimMetrics tools that using Euclidean Distance algorithm. Then, two scores obtained from matched concepts and matched relations are weighted together to form a “total

TABLE 1
THE RESULT OF THE SIMILARITY CALCULATION ODPS

No	ODPs Type	Name	Similarity Value
1	Domain_Modeling ODP	Adapted_SEP	52%
2		Composite Property Chain	62%
3		Interactor Role Interaction	52%
4		List	46%
5		Sequence	51%
6	Extension ODP	Exception	42%
7		Nary Data Type Relationship	52%
8		Nary Relationship	54%
9	Good Practice	Closure	71%
10		Defined Class Description	56%
11		Entity Feature Value	47%
12		Entity Property Quality	56%
13		Entity Quality	55%
14		Normalization	50%
15		Selector	38%
16		Upper Level Ontology	17%
17		Value Partition	44%

matching-score" for each pattern. Then a decision is made according to some threshold value, the patterns will be kept and included in the ontology result, which will be discarded. Finally, an ontology is built from the accepted patterns that has the highest score similarity.

Score Computation

At this stage similarity calculation are computed between the extracted term and relation of the concepts and the relationships that exist in the design pattern. This stage using tools called SimMetrics. In this tool, there are various algorithms for example Euclidean Distance similarity distance, Levenshtein, and others. At this stage, average values are calculated from all the existing algorithms, so can be obtained value or score for string matching. The result in the score computation stage is the value or similarity score for each design pattern. Afterwards, a design pattern that has the highest similarity score is implementted to build ontology. We give more attention for relation between the concept because it can make ontology more structured.

Ontology Building

Ontology building is the stage to build an ontology of terms and relations that correspond to ontology design patterns. Ontology which implements a design pattern that has highest similarity value and Alzheimer ontology has been built on a previous study [5]. This stage uses tools to produce named OWL ontology generation to build ontology from terms and relationships that exist.

The step to using ontology generation is the first we must search definition of the term which entered. The search is connecting with PubMed in the protégé. After that, the automatic mapping of terms and relation that exist as to build a new ontology.

Ontology Evaluation

Ontology evaluation can be viewed in terms of complexity, time and effort required to build this ontology. This evaluation compared with the result by Drame et al [5] that construct semi-automatic ontology. Moreover, ontology evaluation also calculates accuracy of the terms and relation that used to build the ontology. Accuracy is calculated by equation(1).

$$\boldsymbol{accuracy = x/y} \quad (1)$$

Where, x is matching results of term/relation and y is total all of match term/relation.

The meaning of matching results of term or relation is match terms and relations that extracted from corpus with the terms and relations in design patterns that have the highest score similarity.

The meaning of total all term/relation is all of the terms and relations that extracted from corpus and has been filtered by Alzheimer's Glossary. That terms and relations are matched with the terms and relation on ODPs.

3. Results and Analysis

Term & Relation Extraction

The corpus used in this research includes 125 papers. The results of term and relation extraction are 1995 terms and 42 relations between terms. The number of terms resulting from the extraction of corpus was very large, so it needs to filtering terms that have association with Alzheimer's disease.

Matching with Glossary Alzheimer

At this stage, the result of term and relation extraction is the filtering with a matching Alzheimer's glossary. Alzheimer's glossary contains 370 terms related to Alzheimer. Once matched, the term acquired a number of 350 terms. This is different from the terms extracted from a corpus

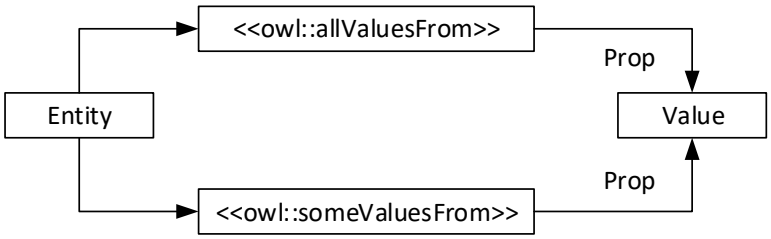


Figure 2. Structure of ODPs closure

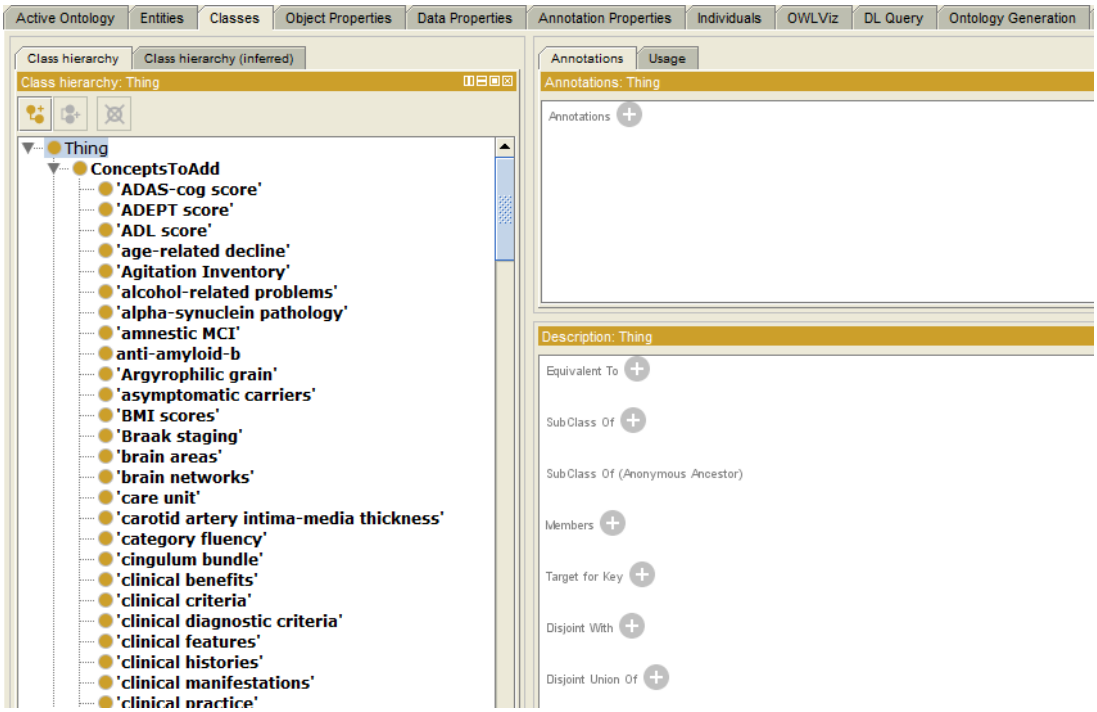


Figure 3. Visualization one part of the ontology in protégé

using extraction with text2onto because the extraction term will be related many health term in general, not specifically related to Alzheimer's disease. In addition, the number of terms in the Alzheimer's glossary of much less than terms of any health glossary in general so the scope of term filtering will be limited.

Matching with Ontology Design Patterns

Term and relation that has been filtered will be matched with a list of terms and relation that exist in the ontology design patterns. In the catalog there are several kinds of ontology design patterns (ODPs). The matching results will then be calculated for the similarity values between terms and filtering results with a term relation and relation that exist in the ontology design patterns (ODPs).

Score Computation

The result of similarity matching between term and relation with each ontology design patterns is shown in table 1.

The highest value of similarity found in ontology design patterns closure is equal to 71%. Closure ontology design pattern is a design pattern that limits the relationships among concepts which allows it to happen by clarifying the relation [11]. The limitations in this relation allows to express a concept has had a particular relation and only those relation. For example, a carnivorous is a meat eating animals, with closure design pattern can be revealed that carnivores do not eat other foods besides meat.

Ontology Building

Fully automatic ontology in this research consists of several components, there are 381 terms and

184 relations. Terms and relations are used to build ontology with tools namely OWL ontology generation. The figure 3 represents the results of the ontology that has been built in the protégé editor tool. There are 200 new terms and 42 new relations were added in that ontology.

The method of this ontology construction can be applied to other domains using a bilingual corpus associated with that domain and use the glossary or dictionary associated with that domain to filtering the terms and relations associated with that domain. If there is a bilingual corpus and glossary related to a specific domain ontology that can be built using the stage as a method of this research

Ontology Evaluation

Ontology evaluation can be viewed in terms of complexity, time and effort required to build this ontology. The result of evaluation is fully automatic ontology construction that can shorten the development time compared to ontology manually or semi-automatic which requires expert validation for a month. In previous studies it takes two teams in the field of Alzheimer's expert to validate the built of ontology. New term and relation in fully automatic ontology construction present that the ontology more complexity than semi automatic ontology in previous research [5].

The result of accuracy value of fully automatic ontology construction is 72%. It is obtained from the calculation of the number of terms or relations corresponding number of 525 terms or relations and the total term or relation in the ontology built a number of 726 terms or relations. This indicates that fully automatic ontology construction method used in the study was quite nice to be able to build the ontology, but it still needs to be improved in order to obtain higher accuracy values.

This accuracy value can not be general in this research, because the accuracy value can be different for other cases. However, the accuracy value can be as a supporting material to the evaluation of this research.

4. Conclusion

This research succeeds to make fully automatic bilingual domain ontology using the Ontology Design Patterns (ODPs) and text corpora. The result of ontology development includes 381 terms and 184 relations with addition of 200 terms and 42 new relations.

Fully automatic construction could speed up and reduce the human's role as expert to evaluate ontology rather than building ontology manually. The result of evaluation is fully automatic ontology construction that can shorten development time compared to manual ontology or semi-automatic which requires expert validation. New term and relation in automatic ontology construction present that the ontology are more complicated than semi automatic ontology in previous research.

For future work, addition number of term in Alzheimer's glossary is recommended to filter the term results of a corpus extraction well. Alzheimer's glossary can improve the results of filter term from extraction corpus. In addition, type of data ontology design patterns (ODPs) can be improved to get the highest similarity value for selected design patterns that will be implemented to build ontology.

Moreover, ontology enrichment to increase the number of terms can be implemented in ontology building. Ontology enrichment using parallel corpora of the website in English and French can obtain terms and synonymous terms in other languages. This process uses term alignment with alignment approach to enrich bilingual biomedical resource. After that, parallel term can integrate into ontology that has been built.

Fully automatic ontology construction not only can be applied to Alzheimer's domain knowledge, but also fully automatic ontology construction methods can be applied to other domain knowledge. Therefore, ontology construction on other domain ontology can be used for research and development of their domain knowledge.

References

[1] Louis, Jean L. Prototype System For Automatic Ontology Construction. Thesis Magister Information Technology. The Royal Institute Of Technology. Sweden. 2007.

[2] Gruber TR. A translation approach to portable ontology specifications. Knowledge Acquisit 1993;5:199–220.

[3] J. Cimiano, Philipp and Völker, "A Framework for Ontology Learning and Data-Driven Change Discovery," Nat. Lang. Process. Inf. Syst., pp. 227– 238, 2005.

[4] Blomqvist, E. Fully Automatic Construction of Enterprise Ontologies Using Design Patterns: Initial Method and First Experiences. In Proceedings of OTM 2005 Conferences, Ontologies, DataBases, and Applications of Semantics (ODBASE), Agia Napa, Cyprus, Oct 31- Nov 4, 2005.

[5] Dramé K et al. Reuse of termino-ontological resources and text corpora for building a multilingual domain ontology: An application

to Alzheimer's disease. J Biomed. Inform., vol. 48, pp. 171–182, 2014.

[6] Dahab, M.Y., Hassan, H. and Rafea, A., "TextOntoEx: Automatic ontology construction from natural English text," Expert Syst. Appl., vol. 34, pp. 1474–1480, 2008.

[7] Chen, R. C., Liang, J. Y., and Pan, R. H., "Using recursive ART network to construction domain ontology based on term frequency and inverse document frequency," Expert Syst. Appl., vol. 34, pp. 488–501, 2008.

[8] Navigli, R., and elardi, P., "From Glossaries to Ontologies : Extracting Semantic Structure from Textual Definitions," Ontol. Learn. Popul. Bridg. Gap between Text Knowl., pp. 71–87, 2008.

[9] Chapman, S., Norton, B., and Ciravegna, F., "Armadillo: Integrating knowledge for the semantic web," Proc. Dagstuhl Semin. Mach. Learn. Semant. Web, pp. 2–4, 2005.

[10] Wächter, T., and Schroeder, M., "Semi-automated ontology generation within OBO-Edit," Bioinformatics, vol. 26, pp. 88–96, 2010.

[11] ODP public catalog. Closure. http://www.gong.manchester.ac.uk/odp/html/Closure html. Access on Monday, 26 May 2014. 09.00 am.

Jurnal Ilmu Komputer dan Informasi (Journal of a Science and Information). 10/2 (2017), 67-76
DOI: http://dx.doi.org/10.21609/jiki.v10i2.375

RANDOM ADJUSTMENT - BASED CHAOTIC METAHEURISTIC ALGORITHMS FOR IMAGE CONTRAST ENHANCEMENT

Vina Ayumi, L. M. Rasdi Rere, Mochamad I. Fanany, and Aniati M. Arymurthy

Faculty of Computer Science, Universitas Indonesia, Kampus UI, Depok, 16424, Indonesia

E-mail: vina.ayumi@ui.ac.id, aniati@cs.ui.ac.id

Abstract

Metaheuristic algorithm is a powerful optimization method, in which it can solve problems by exploring the ordinarily large solution search space of these instances, that are believed to be hard in general. However, the performances of these algorithms significantly depend on the setting of their parameter, while is not easy to set them accurately as well as completely relying on the problem's characteristic. To fine-tune the parameters automatically, many methods have been proposed to address this challenge, including fuzzy logic, chaos, random adjustment and others. All of these methods for many years have been developed indepen- dently for automatic setting of metaheuristic parameters, and integration of two or more of these methods has not yet much conducted. Thus, a method that provides advantage from combining chaos and random adjustment is proposed. Some popular metaheuristic algo- rithms are used to test the performance of the proposed method, i.e. simulated annealing, particle swarm optimization, differential evolution, and harmony search. As a case study of this research is contrast enhancement for images of Cameraman, Lena, Boat and Rice. In general, the simulation results show that the proposed methods are better than the original metaheuristic, chaotic metaheuristic, and metaheuristic by random adjustment.

Keywords: *metaheuristic, chaos, random adjustment, image contrast enhancement*

Abstrak

Algoritma Metaheuristic adalah metode pengoptimalan yang hebat, di mana ia dapat memecahkan masalah dengan menjelajahi ruang pencarian solusi yang biasanya besar dari contoh-contoh ini, yang diyakini sulit dilakukan secara umum. Namun, kinerja algoritme ini sangat bergantung pada pengaturan parameter mereka, namun tidak mudah untuk menetapkannya secara akurat serta sepenuhnya bergantung pada karakteristik masalah. Untuk menyempurnakan parameter secara otomatis, banyak metode telah diajukan untuk mengatasi tantangan ini, termasuk logika fuzzy, kekacauan, penyesuaian acak dan lain-lain. Semua metode ini selama bertahun-tahun telah dikembangkan secara terpisah untuk penentuan parameter metaheuristik secara otomatis, dan integrasi dua atau lebih dari metode ini belum banyak dilakukan. Dengan demikian, metode yang memberikan keuntungan dari penggabungan kekacauan dan penyesuaian acak pun diusulkan. Beberapa algoritma metaheuristik populer digunakan untuk menguji kinerja metode yang diusulkan, yaitu simulasi anil, optimasi partikel, evolusi diferensial, dan pencarian harmonis. Sebagai studi kasus penelitian ini adalah peningkatan kontras untuk citra Cameraman, Lena, Boat and Rice. Secara umum, hasil simulasi menunjukkan bahwa metode yang diusulkan lebih baik daripada metaheuristik asli, metaheuristik kacau, dan metaheuristik dengan penyesuaian acak.

Kata Kunci: *metaheuristik, chaos, penyesuaian acak, peningkatan kontras gambar*

1. Introduction

Image enhancement is one of the main concerns in image processing that aims to improve the appearance of an image, to enhance their visual quality on human eyes, including to sharpen the features and to increase the contrast. Image enhancement it is useful to further image application, such as facilitating image segmentation, recognizing and interpreting useful information from the image, but does not increase nor decrease the essential information of the original image. In general, the image enhancement methods can be divided into four classes, i.e. contrast enhancement, edge enhancement, noise enhancement and edge restoration [1]. Among those techniques, contrast enhancement is the focus of this paper.

There are many variations of image enhancement algorithms have been proposed. Some of the famous methods are contrast manipulations and histogram equalization for enhancing the contrast image. Contrast manipulations or linear

contrast stretching employs a linear transformation that remaps the gray-levels in a given image to fill the full range of values, and histogram equalization applies a transformation that produces a close to uniform histogram for the relative frequency of the gray-levels in the image [2].

In recent times, many metaheuristic methods have been developed for image processing applications, including image enhancement problems. Some paper [2-4] report that these methods outperform for image contrast enhancement than classical point operation. Based on some principles of biology, physics or ethology; almost all of metaheuristic are nature-inspired. Other classifications form of this method is single-solution and population-based based metaheuristic [5].

Three main purposes of metaheuristic algorithm: solving large problems, solving prob- lems faster, and obtaining robust algorithms [6]. Besides, they are simple to design, flexible, and also not difficult to implement. However, setting parameters of these methods are not easy, and entirely depend on the problems. Some of the methods have been recommended to adjust the parameters of metaheuristic automatically . Liu and Lampinen [7] proposed FADE (fuzzy adaptive differential evolution), where the fuzzy logic is used to ad the param- eter controls of mutation and crossover. Di and Wang [8] use harmony search with chaos for training RBFNN (radial basis function neural network). Coelho et al. [1] use chaos to optimize DE for image contrast enhancement. Ferens et al [9] proposed CSA (chaotic simulated annealing) for task allocation in a multiprocessing system. Noman et al. [10] proposed adaptive DE (aDE) based on random adjustment, where the strategy is by com- paring the objective of spring with the average value of the current generation. Li et al. [11] introduced market-oriented task-level scheduling in cloud workflow systems using chaos to particle swarm optimization (PSO).

All of that methods have each of advantage on automatically adjusting of metaheuristic parameters, however, integration two or more of them are rarely conducted. In this paper, we integrate 2 methods, chaos and random adjustment for attaining benefit from both of them. Chaos can be used to avoid being trapped into a local minimum and to enrich the searching behavior. On the contrary, random adjustment can be applied to achieve greater accuracy. Four types of metaheuristic algorithms are selected to represent all categories for test the proposed method performance: physics phenomena and a single solution based represent by SA, biology phenomena and population-based represent by DE, ethologic phenomena and also population-based represent by PSO, and musical phenomena as well as population-based, represents by HS.

This paper is organized as follows: Section 1 is introduction; Section 2 gives description of image contras enhancement; Section 3 describe the proposed methods; Section 4 we present simulation result; and conclusion of this paper in Section 5 (one blank single space line, 10 pt)

2. Methods

Image Contrast Enhancement

Contrast enhancement is applied to transform an image based on the psychophysical characteristics of the human visual system. Two techniques that are usually used for contrast enhancement are indirect and direct methods of contrast enhancement [12]. The indirect image contrast enhancement algorithms enhance the image without measuring the contrast. The direct local contrast enhancement algorithms create a criterion of contrast measure and improving the contrast measurement directly to enhance the images [1]. The proposed method in this paper are applied using a direct image enhancement approach to adjust the gray-level intensity transformation in the image. The setting up of a suitable image contrast measure is a critical step in direct image enhancement approach.

In spatial domain to the gray-level image, the enhancement uses transformation function. To generate the enhanced image, the transformation function generates a new intensity value for each pixel of original image as shown in equation(1).

$$h(i,j) = T[f(i,j)] \quad (1)$$

where f(i, j) is the gray value of the (i, j)th pixel of the input images, h(i, j) is the gray value of the (i, j)th pixel of the enhanced images and S is the transformation function [4].

The contrast of the image can be measured locally and globally. A local contrast functions regarding the relative difference between a central region and a larger surrounding area of a given pixel. By some of the contrast enhancement functions, the contrast values are then enhanced. The enhancement function such as the square root function, the exponential, the logarithm and the trigonometric functions [12].

The transformation functions T that is based on the gray-level distribution in the neighborhood of every pixel in a given image, applied by local enhancement methods [7]. The following method applies to each pixel at the location (x, y) shown in equation (2), is used in this paper for a transformation function as shown in equation(2).

$$T[f(i,j)] = \left(s\frac{M}{\sigma(x,y)+q}\right).[f(x,y) - r.m(x,y)] + m(x,y)^p \quad (2)$$

where (x,y) and m(x, y) are the standard deviation and the gray-level mean respectively, computed in a neighborhood centered at (x, y). Where M is the global mean of the image, f(x, y) and g (x, y) is the gray-level intensity and the pixels output gray-level intensity value of input image pixel at location (x, y) [1].

A nonzero value for q in (2) allows for zero standard deviation in the neighborhood while c allows for only a fraction of the mean m(x, y) to subtracted from the original pixels gray-level f(x, y). The last term m(x, y)p may have a brightening and smoothing effect on the image. The parameters of p, q, r and s defined over the positive real number and they are the same for the whole image [3]. According to an objective function that describes the contrast of the image, the task of metaheuristic in this formula is to find the combination of parameters p, q, r and s.

A criterion for enhancement method should be chosen to apply an automatic image enhancement technique, which does not require human intervention and no objective parameters are given by the user. This criterion will be directly related to the objective function of the metaheuristic methods. The objective function adopted in this paper for an enhancement criterion shown in equation(3).

$$F(M) = \mathbf{log}\left(\mathbf{log}\left(E\left(I(M)\right)\right)\right).\frac{ne\left(I(M)\right)}{PH.PV}. \quad (3)$$
$$H(I(M))$$

Function F(M) and I(M) denote an objective function for maximization problem and the original image I with the transformation T in each pixel at location (x, y) applied according to Eq. (1). Where the respective parameters p, q, r, and s are given by the M = (p q r s). Furthermore, E(I(M)) is the intensity of the edges detected with a Sobel edge detector that is applied to the transformed image I(M). ne(I(M)) is the number of edge pixels as detected with the Sobel edge detector, PH and PV are the number of pixels in the horizontal and vertical direction of the image, respectively. Lastly, the entropy of the image I(M) measured by H(I(M)) [1].

Proposed Method

Most of the metaheuristic algorithms have relevant parameters, such as amplification factor (F) and crossover rate (CR) in DE, initialize temperature (T) and reduction factor (c) in SA, harmony memory considering rate (HMCR) and pitch adjusting rate (PAR) in HS, as well as acceleration coefficients (c1, c2) in PSO. All of these parameters are usually sensitive, in while an improper setting of them can result in the poor performance of the system. Some studies have been conducted to adjust automatically these parameters based on the characteristic of the problems, including fuzzy logic, chaos, random adjustment, and others.

In this paper, we proposed a combination of chaos and random adjustment to improve the performance of some metaheuristic algorithms. Characteristic of chaos is nonlinear systems. Although it looks like to be stochastic, then it occurs in a deterministic nonlinear system under deterministic condition [13]. This method can avoid being trapped into local optimum and improve the performance of searching [1]. One of the systems is chaotic sequence, defined in equation(4).

$$x(n) = \mu.x(n-1).[1-x(n-1)] \quad (4)$$

where n and μ is sample parameter and control parameter. Substantially both of the param- eters decides whether x stabilizes at a constant size, behaves chaotically in an unpredictable pattern, or oscillates between a limited sequence of sizes. A very small difference in the initial value of x causes substantial differences in its long-time behavior. In this work, the variety of μ is $1 < \mu < 4$; x is distributed in the range [0, 1] provided the initial $x(1) \notin 0, 0.25, 0.50, 0.75, 1$.

In case of random adjustment, for instances DE algorithms, the strategy by comparing objective value of the offspring $f(x_N^{child})$ with the average of objective value in current generation favg . If $f(x_N^{child})$ is better than f_{avg} , then mutation factor and crossover rate of the primary parent are retained in offspring, or else the parameters are changed randomly.

Random Adjustment-based Chaotic SA

Simulated annealing (SA) is a robust and compact technique was first proposed by Kirk-patrick et al. [14]. With a substantial reduction in computation time, SA provides excellent solutions to single and multiple objective optimization problems. The origin of this method is Metropolis algorithm [15]. Inspired by annealing technique, this method aims to obtain the solid state of minimal energy or ground states of matter. This technique consists in heating a material to the high temperature, then in lowering the temperature slowly.

The Boltzmann distribution is the quantitative key of SA method which species that the

probability of being in any particular state x is given by equation(5).

$$p(x) = e^{\frac{-\Delta f(x)}{kT}} \quad (5)$$

where f(x) is the energy of the configuration, k is Boltzmanns constant, and T is temperature. In this paper, we proposed 3 variant of methods for chaotic SA based on random adjustment. First is CSARA-1, where parameter of k is replaced by generating the value from chaotic sequence. Otherwise, the reduction factor parameter c is adjusted randomly. This value of c is used in a process when the result of the new objective function is better than the old objective function, or when the random value r is bigger than the Boltzmann distribution p(x). This process will continue until the desired criteria have been achieved.

The second variant is CSARA-2, by replacing parameter of c with chaotic sequence and parameter of k is selected randomly. As long as the new objective function is better than the old objective function, or the random value of r is bigger than the value of p(x), the value of k is still used in the process.

The third variant is CSARA-3, in which the parameter of k is constant, and c is produced from a chaotic sequence. The value of c is not substituted, as long as the new objective function is better than the old objective function, or the random value of r is bigger than the value of p(x).

Random Adjustment-based Chaotic DE

Differential Evolution (DE) is one of the latest evolutionary algorithms proposed by Price and Storn in 1995 that applied to a continuous optimization problem. This method proposed to solve the chebyshev polynomial fitting problem and have proven for many different tasks to be a very reliable optimization strategy [5]. Starts by sampling the search space at multiple, DE algorithm randomly selected search points and creates new search points through perturbation of the existing points. DE creates new search points which are evaluated against their parents using the operation of differential mutation and recombination. Furthermore to promote the winners to the next generation, a selection mechanism is applied. Until the termination criterion is satisfied, this cycle is iterated [10]. Price et al. have suggested different variant of DE, which are conventionally named DE/x/y/z. DE/rand/1/bin is the classical version as shown in equation(6), the target vector is randomly selected in mutation process, and only one different vector is used. The acronym of bin indicates a binomial decision rule that controlled the crossover.

$$x_G^{mut} = x_G^{r1} + F(x_G^{r2} - x_G^{r3}) \quad (6)$$

In this paper, we proposed 3 variant methods for DE. First is CDERA-1, where CR parameter is generated by chaotic sequence and mutation factor F is created randomly. On condition that new objective function is better than the average of old objective function, parameter of F is kept in used in the process. However, if not the new parameter of F is created randomly. All of the procedure will continue until the termination criterion is satisfied. The second variant is CDERA-2, where F parameter is created by chaotic sequence and CR parameter is selected randomly. In case of the new objective function is better than the average of old objective function, CR is kept in used in the process. Otherwise, CR is created randomly. The third variant is CDERA-3, in which the parameter of F is constant, and CR is created from chaotic sequence. The value of CR is not replaced, as long as the new objective function is better than the average of old objective function. Otherwise, it uses the next value of chaotic sequence.

Random Adjustment-based Chaotic PSO

Particle swarm optimization (PSO) is an adaptive algorithm based on social-psychological metaphor; a population of particles adapts by returning stochastically toward previously successful regions. The metaphor of the flocking behavior of birds uses by PSO to solve an optimization problem. Introduced in 1995 by J. Kennedy and R. Eberhart [16], this method was initial as a global optimization technique. In this algorithm, many particles are stochastically generated in the search space. As a candidate solution to the problem, each particle is represented by a velocity, a location in the search space and has a memory which helps it in remembering its previous best position. In the initialization phase of PSO, the position and velocities of all individuals are randomly initialized. The velocity defines direction and distance of particle should go. It is updated according to the equation(7).

$$v_i^{j+1} = wv_i^j + c_1 r_1.\left[p_{ibest} - x_i^j\right] + c_2 r_2.\left[g_{best} - x_i^j\right] \quad (7)$$

where i = 1, 2, , N. N is the size of the swarm; p_{ibest} is the particle best-reached solution and g_{best} is the swarm global best solution. Two random numbers r1 and r2 are uniformly distributed in the

range [0,1], constant multiplier terms c1 and c2 are known as acceleration coefficients. They represent the attraction that a particle has either towards its own success or towards the success of its neighbors, respectively.

To overcome the premature convergence problem of PSO, the inertia weight ω is used. A large inertia weight encourages global exploration while a smaller inertia weight encourages local exploitation [5]. The position of each particle x_i^j is also updated in every each iteration by adding the velocity vector v_i^{j+1} to the position vector, using equation(8).

$$x_i^{j+1} = x_i^j + v_i^{j+1} \quad (8)$$

In this paper, we proposed three alternative methods for PSO. First is CPSORA-1, where parameters of r1 and r2 are replaced by generating their values from the chaotic sequence. These values of r1 and r2 are kept in used as long as the new objective function is better than average objective function. Otherwise, the next value of the chaotic sequence is used. This process will continue until the desired criteria have been achieved. The second variant is CPSORA-2, where r1 is a constant value, and r2 is created from the chaotic sequence. On condition that the new objective function is better than the average of old objective function, r2 is kept in used in the process. Otherwise, it uses the next value of chaotic sequence The third variant is CPSORA-3, where essentially is the same with the second variant, but in this case, r1 is created from the chaotic sequence, and r2 is a constant value.

Random Adjustment-based Chaotic HS

Harmony search (HS) proposed by Zong Woo Geem et al in 2001 is a search algorithm considered to be a population-based. By the musical process of searching for a perfect state harmony, this method is inspired. The optimization solution vector analogous to the harmony in music, and the local and global search schemes in optimization techniques analogous to the musicians improvisations. The HS algorithm uses a stochastic random search that is based on the harmony memory considering rate (HMCR) and the pitch adjusting rate (PAR) so that derivative information is unnecessary [17].

Three possible options exist when a musician improvises one pitch: (1) playing any one pitch from his/her memory, (2) playing an adjacent pitch of one pitch from his/her memory, (3) playing a totally random pitch from the possible sound range. Similarly, when each decision variable chooses one value in the HS algorithm, it follows any one of three rules: (1) choosing any one value from HS memory (defined as memory considerations), (2) selecting an adjacent value of one value from the range (defined as pitch adjustments), (3) choosing the random value from the possible value range (defined as randomization)[17].

In this paper, 3 alternative methods for HS are proposed. First is CHSRA-1, where parameters of HMCR and PAR are replaced by generating their values from the chaotic sequence. These values are kept in used, as long as the new objective function is better than the averages the old objective function. Otherwise, the next value of the chaotic sequence is used. This process will continue until the desired criteria have been achieved. The second variant is CHSRA-2, where HMCR parameter is created by chaotic sequence and PAR parameter is selected randomly. In case of new objective function is better than average objective function; PAR is kept in used in process, else PAR is created randomly. The third variant is CHSRA-3, wherein essential is the same with the second variant. In this case, PAR is created from chaotic sequence, and HMCR is selected randomly.

3. Results and Analysis

The optimization problem in this paper is to enhance the image contrast using chaotic metaheuristic algorithms based on random adjustment approaches. The simulation objective is to increase the overall intensity at the edges, increasing the measurement of entropy, and maximize the number of pixel in the edges. Moreover, the simulation with original metaheuristic, metaheuristic use chaotic sequence, and metaheuristic by random adjustment are also conducted.

Since to ensure the control parameters in metaheuristic is difficult, we decided to run 30 times for all images in all simulation, as well as for all the methods stopping criterion is 40. We also set all the parameters that are looked for as M = (p q r s), with boundaries: p = [0 1.5], q = [0 2], r = [0.5 2], and s = [0.5 30].

TABLE 1
SIMULATION RESULTS OF SA, SARA, CSA, AND CSARA

Methods	Lena		Boat		Cameraman		Rice	
	M1	T1	M2	T2	M3	T3	M4	T4
SA1	0.1515	43.17	0.1323	36.89	0.1480	47.73	0.2494	71.83
SA2	0.1542	119.83	0.1346	78.48	0.1539	103.36	0.2500	104.44
SA3	0.1537	286.30	0.1342	212.42	0.1572	174.46	0.2494	317.36
SARA1	0.1544	259.12	0.1328	179.23	0.1545	239.78	0.2497	182.84
SARA2	0.1534	99.76	0.1330	75.64	0.1529	70.26	0.2508	68.43
SARA3	0.1544	232.97	0.1346	177.40	0.1589	238.57	0.2510	183.42
CSA1	0.1530	208.42	0.1323	181.72	0.1580	227.80	0.2510	187.95
CSA2	0.1545	89.19	0.1350	172.80	0.1527	94.58	0.2503	97.02
CSA3	0.1531	271.93	0.1336	227.02	0.1514	185.21	0.2500	240.01
CSARA1	0.1549	210.74	0.1343	206.11	0.1521	187.06	0.2510	183.05
CSARA2	0.1551	207.87	0.1351	180.32	0.1568	242.06	0.2508	121.18
CSARA3	0.1541	259.45	0.1331	175.00	0.1508	239.89	0.2510	191.98

TABLE 2
SIMULATION RESULTS OF DE, DERA, CDE AND CDERA

Methods	Lena		Boat		Cameraman		Rice	
	M1	T1	M2	T2	M3	T3	M4	T4
DE1	0.1577	134.83	0.1343	112.89	0.1587	120.27	0.2536	117.23
DE2	0.1568	137.30	0.1412	87.54	0.1576	119.43	0.2529	118.70
DE3	0.1513	140.33	0.1310	113.39	0.1515	120.35	0.2476	118.37
DE4	0.1504	136.90	0.1344	113.88	0.1595	120.19	0.2481	115.91
DE5	0.1517	134.06	0.1352	111.89	0.1507	120.41	0.2477	91.82
DERA1	0.1549	134.10	0.1399	119.41	0.1562	120.68	0.2531	118.97
DERA2	0.1558	134.06	0.1400	115.87	0.1578	91.21	0.2515	119.50
DERA3	0.1565	134.82	0.1272	115.72	0.1580	116.59	0.2529	119.17
CDE1	0.1559	105.83	0.1418	86.41	0.1578	106.46	0.2529	91.67
CDE2	0.1568	134.63	0.1406	115.54	0.1582	120.70	0.2531	119.04
CDE3	0.1532	133.89	0.1413	113.29	0.1580	91.04	0.2530	118.46
CDERA1	0.1558	104.42	0.1398	87.46	0.1575	91.22	0.2529	88.26

All of the algorithms were programmed and implemented in MatlabR211a, on personal computer with processor Intel Core i7-4500U, 8 GB RAM running memory, in Windows 8.1. To evaluate the image enhancements based on these proposed methods, four images were evaluated, i.e. Cameraman, Rice, Boat, and Lena; all of them have been resized at 256x256 pixels, and are converted into double precision for numerical computation. In case of contrast color image enhancement, at the first time, the RGB color spaces (red, green, blue) is converted into YIQ color space (luminance, hue, saturation), and then apply them to the methods only for the Q component. After that process, they convert back to the RGB color space

Simulation of SA, SARA, CSA and CSARA

Simulation of simulated annealing algorithm is carried out in 13 conditions. First, group is 3 simulations on the original of simulated annealing: SA1 (k=1,c=0.2), SA2 (k=1,c=0.5), SA3 (k=1,c=0.8). Second is 4 simulations on SA by random adjustment: SARA1 (k = RA, c = 0.2), SARA2 (k = RA, c = 0.5), SARA3 (k = RA, c = 0.8), SARA4 (k = c = RA). Third is 3 simulations on SA by chaotic: CSA1 (k = Ch, c = 0.2), CSA2 (k = Ch, c = 0.5), CSA3 (k = Ch, c = 0.8), and fourth is 3 simulations on the proposed methods, i.e. chaotic SA based on random adjustment: CSARA1 (k = Ch, c = RA), CSARA2 (k = RA, c = Ch), CSARA3 (k = 1, c = ChRA).

Simulation results for all SA algorithms are given in Table 1. These results show that mean objective function of the proposed methods achieves the higher value for all images: CSARA3 for image of Lena (M1 = 0.1554), CSARA1 for image of Boat (M2 = 0.1351), Cameraman (M3 = 0.1584) and rice (M4 = 0.2512). In case of computation time, the comparisons from the simulation results for SA algorithms shows that for all images, the best computa- tion time is SA1: Lena (T1 = 43.17s), Boat (T2 = 47.73s), Cameraman (T3 = 36.90s), , and Rice (T4 = 71.83s).

Moreover, the best objective function of proposed methods also gives better value for all images: CSARA1 with SARA3 for the image of

TABLE 3
SIMULATION RESULTS OF PSO, PSORA, CPSO AND CPSORA

Methods	Lena		Boat		Cameraman		Rice	
	M1	T1	M2	T2	M3	T3	M4	T4
PSO1	0.1525	131.11	0.1361	111.94	0.1523	118.28	0.2509	116.52
PSO2	0.1515	132.71	0.1357	112.51	0.1505	115.84	0.2521	86.85
PSO3	0.1506	131.15	0.1333	113.85	0.1438	94.20	0.2508	86.86
PSO4	0.1508	133.16	0.1339	113.57	0.1527	98.16	0.2510	86.90
PSORA1	0.1533	107.80	0.1379	92.72	0.1557	95.61	0.2522	88.05
PSORA2	0.1550	104.60	0.1382	87.28	0.1559	91.54	0.2518	90.65
PSORA3	0.1539	100.81	0.1353	85.20	0.1544	90.20	0.2513	91.17
CPSO1	0.1512	105.62	0.1368	88.73	0.1572	92.60	0.2521	85.00
CPSO2	0.1549	103.26	0.1372	85.65	0.1573	92.05	0.2526	88.49
CPSO3	0.1513	101.99	0.1360	84.68	0.1525	89.31	0.2517	88.61
CPSORA1	0.1548	139.54	0.1387	120.00	0.1580	125.49	0.2524	124.54
CPSORA2	0.1557	101.07	0.1378	111.93	0.1586	89.26	0.2529	91.22

TABLE 4
SIMULATION RESULTS OF HS, HSRA, CHS AND CHSRA

Methods	Lena		Boat		Cameraman		Rice	
	M1	T1	M2	T2	M3	T3	M4	T4
HS1	0.1409	7.84	0.1298	6.32	0.1397	8.89	0.2354	6.68
HS2	0.1428	7.71	0.1297	6.37	0.1391	8.95	0.2359	8.68
HS3	0.1383	7.69	0.1290	6.49	0.1398	9.01	0.2418	14.47
HS4	0.1403	7.77	0.1245	6.59	0.1378	8.94	0.2355	10.28
HS5	0.1441	7.60	0.1279	6.53	0.1402	8.97	0.2385	6.70
HSRA1	0.1441	10.00	0.1287	8.65	0.1405	8.92	0.2407	8.88
HSRA2	0.1438	9.95	0.1272	6.44	0.1408	8.79	0.2403	6.66
HSRA3	0.1448	9.90	0.1303	8.29	0.1388	8.79	0.2421	6.63
CHS1	0.1440	9.81	0.1290	6.38	0.1403	8.93	0.2419	6.76
CHS2	0.1457	9.92	0.1294	8.41	0.1384	6.65	0.2423	8.63
CHS3	0.1441	7.56	0.1304	7.42	0.1384	6.67	0.2423	6.52
CHSRA1	0.1460	7.77	0.1301	6.38	0.1416	6.94	0.2424	6.86

Lena (0.1590); CSARA1 with SARA3, SARA4, and CSA2 for image of Boat (0.1436); CSARA3 for images of Cameraman (0.1703) and Rice (0.2539). In case of the worst objective function, the original of SA gives the less value for all images: SA1 for images of Lena (0.1301) and Cameraman (0.1154); SA2 for the image of Rice (0.2278); SA3 for the of Boat (0.1162).

Simulation of DE, DERA, CDE and CDERA

Simulation of differential evolution algorithm is carried out in 14 conditions. First, group is 5 simulations on the original differential evolution: DE1 (F = CR = 0.8), DE2 (F = CR = 0.5), DE3 (F = CR = 0.2), DE4 (F = 0.8, CR = 0.2), DE5 (F = 0.2, CR = 0.8). Second is 3 simulations on DE by random adjustment: DERA1 (F = CR = RA), DERA2 (F = RA, CR = 0.5), DERA3 (F = 0.5, CR = RA). Third is 3 simulations on chaotic DE: CDE1 (F = CR = Ch), DE2 (F = Ch, CR = 0.8), DE3 (F = 0.8, CR = Ch) and fourth is 3 simulations on the proposed methods, i.e. chaotic DE based on random adjustment: CDERA1 (F = RA, CR = Ch), CDERA2 (F = Ch, CR = RA), CDERA3 (F = 1, CR = ChRA).

Simulation results for all DE algorithms are given in Table 2. These results show that, mean objective function of the proposed methods achieve the higher value only for 2 images: CDERA3 for image of Boat (M2=0.1420) and CDERA2 with DE1 for image of Cameraman (M3=0.1587). Other images are achieved for the higher value of mean objective function on DE1 for images of Lena (M1=0.1577) and Rice (0M4=0.2536). In case of computation times show that, the best computation time for Lena image is CDERA1 (T1=104.42s), Boat image is CDE3 (T2=91.04s), Cameraman image is CDE1 (T3=86.41s), and Rice image is CDERA1 (T4=88.26s).

Furthermore, the best objectives function of proposed methods, only for image of Rice (CDERA2, CDERA3 = 0.2538), together with DE1, DERA1, DERA3, CDE2 and CDE3. Other images are achieved for the higher value of the best objective function on DE1 for image of Lena (0.1592), CDE2 for image Boat (0.1436), and DERA2 for image of Cameraman (0.1655). In a

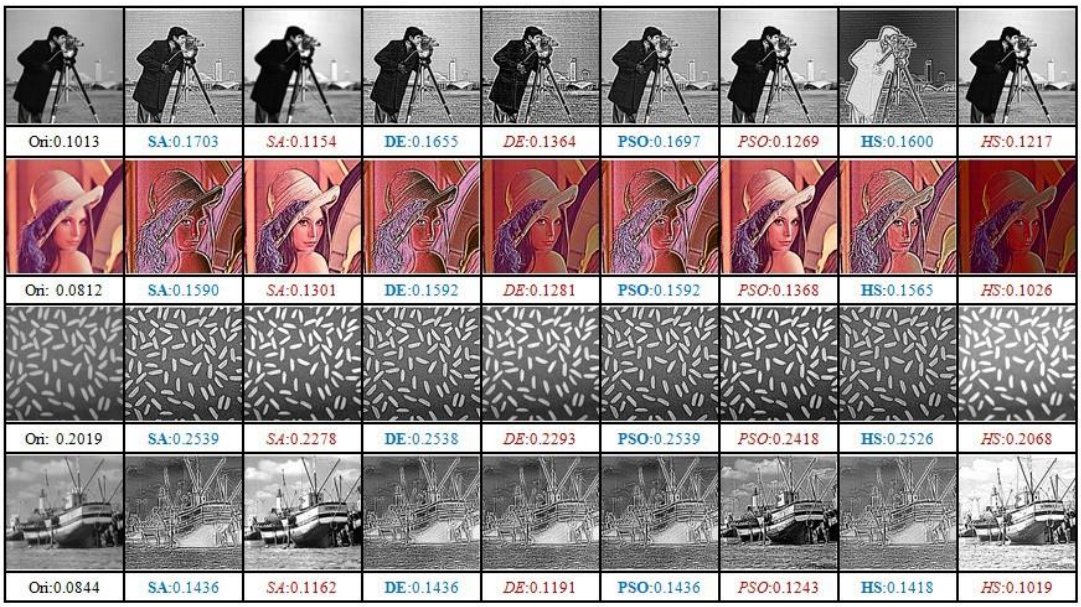


Figure. 1. Comparison of Images for the original (normal and black text), the best objective function (bold and blue text) and the worst objective function (italic and red text) for all metaheuristic algorithms.

case of worst objective function, the original DE gives the less value for all images: DE3 for images of Boat (0.1191), Cameraman (0.1364) and Rice (0.2293) as well as DE5 for image of Lena (0.1281).

Simulation of PSO, PSORA, CPSO and CPSORA

Simulation on particle swarm optimization algorithm is performed in 13 conditions. First group is 4 simulations on the original particle swarm optimization: PSO1 (r1 = r2 = 1.3), PSO2 (r1 = r2 = 1.0), PSO3 (r1 = r2 = 0.5), PSO4 (r1 = r2 = 0.2). Second is 3 simulations on PSO by random adjustment: PSORA1 (r1 = r2 = RA), PSORA2 (r1 =0.8, r2 = RA), PSORA3 (r1 =RA, r2 = 0.8). Third is 3 simulations on chaotic PSO: CPSO1 (r1 = r2 = Ch), CPSO2 (r1 = 1.3, r2 = Ch), CPSO3 (r1 = Ch, r2 = 1.3), and fourth is 3 simulations on the proposed methods: CPSORA1 (r1 = r2 = ChRA), CPSORA2 (r1 =Ch, r2 = RA), CPSORA3 (r1 = RA, r2 = Ch).

Simulation results for all PSO algorithms are given in Table 3. These results show that, mean objective function of the proposed methods achieve the higher value for all images: CPSORA2 for images of Lena (M1 = 0.1557), CPSORA1 for ima-ge of Boat (M2 = 0.1387), Cameraman (M3 = 0.1586) and Rice (M4 = 0.2529). In case of the comparison of computation times shows that the best computation time for Lena image is PSORA3 (T1 = 100.81s), Boat image is CPSORA3 (T2 = 86.71s), Cameraman image is CPSO3 (T3 = 84.68s), and Rice image is CPSO1 (T4 = 85.00s).

Moreover, the best objective functions of proposed methods give a higher value for 3 images: CPSORA1 for image of Cameraman (0.1697); CPSORA1, CPSORA2, CPSORA3 for images Boat (0.1436) and Rice (0.253). In a case of the worst objective function, the original PSO gives the less value for images of Boat (0.1243) and Cameraman (0.1269); CPSO1 for the image of Lena (0.1368); CPSO3 for image of Rice (0.2418).

Simulation of HS, HSRA, CHS and CHSRA

Simulation of harmony search algorithm is conducted in 14 conditions. First, group is five simulations on the original harmony search: HS1 (H = P = 0.8), HS2 (H = P = 0.5), HS3 (H = P = 0.2), HS4 (H = 0.8, P = 0.2), HS5 (H = 0.2, P = 0.8). Second is 3 simulations on HS by random adjustment, HSRA1 (H = P = RA), HSRA2 (H = RA, P = 0.5), HSRA3 (H = 0.5, P = RA). Third is 3 simulations on chaotic HS: CHS1 (H = P = Ch), CHS2 (H = Ch, P = 0.5), CHS3 (H = 0.5, P = RA), and fourth is 3 simulations on the proposed methods: CHSRA1 (H = P = ChRA), CHSRA2 (H = Ch, P = RA), CHSRA3 (H = RA, P = Ch).

Simulation results for all HS algorithms are given in Table 4. These results show that, mean objective function of the proposed methods achieve the higher value for all images: CHSRA1 for images of Lena (M1=0.1460), Cameraman (M3=0.1416) and Rice (M4=0.2424) as well as CHSRA2 for image of Boat (M2=0.1307). In case

TABLE 5
PARAMETER OF BEST OBJECTIVE FUNCTION

Parameter	Cameraman	Lena	Rice	Boat
p	0.6619	0.0644	0.9837	0.0240
q	0.0297	0.8691	2.0000	1.5732
r	1.0065	1.0623	0.9989	1.1124
s	1.2237	29.9987	22.1640	30.0000
E(I(M))	197.8765	366.4818	551.4762	213.6350
ne(I(M))	4057	3732	4989	3685
H(I(M))	7.6171	6.8348	7.6135	6.9511
F(M)	0.1703	0.1592	0.2539	0.1436

TABLE 6
PARAMETER OF WORST OBJECTIVE FUNCTION

Parameter	Cameraman	Lena	Rice	Boat
p	1.4707	0.9561	1.1610	0.7113
q	0.3764	1.1091	1.9994	1.9278
r	0.5000	1.2690	0.6547	0.9401
s	0.8521	5.7182	7.2831	29.6184
E(I(M))	193.5467	85.9574	339.8672	228.6044
ne(I(M))	2813	3546	4944	3352
H(I(M))	7.4834	6.6190	6.7967	5.3437
F(M)	0.1154	0.1026	0.2068	0.1019

of computation times for HS algorithms show that Lena image is CHS3 (T1=7.56 s), Boat image is CHSRA2 (T2=6.63 s) Cameraman image is HS1 (T3=6.32 s), and Rice image is CHS3 (T4=6.52 s).

Moreover, the best objective function of proposed methods gives higher value only for 2 images, which is CHSRA1 for images of (0.1418) and Cameraman (0.1600). The others are HS2 and HSRA2 for images of Lena (0.1565) as well as HSRA1 for the image of Rice (0.2526). In case of the worst objective function, the original HS gives the less value for 3 images: HS1 for image of Rice (0.2068), HS4 for images of Boat (0.1018) and Cameraman (0.1217). Another is CHS1 for the image of Lena (0.1026).

Comparison of images for the original (Ori), the best objective function and the worst objective function on all algorithms are shown in figure 1. Moreover, some examples for combination of parameters p, q, r and s are presented in relation to objective function F as well as intensity of edge E(I(M)) that is detected by Sobel edge detector, number of edge pixels ne(I(M)) and entropy of the images H(I(M)). Simulation results parameters for the best objective function (BOF) are given in Table 5 and for the worst objective function are shown in Table 6.

4. Conclusion

The objective of these proposed methods has been achieved to enhance the detail and the contrast of images. The indicator from the proposed methods is the objective function are better than the original of images. As an example, the mean objective function of Lena image on CSARA1 is 0.1551, while on the original is 0.0812.

Based on the mean objective functions from simulation results, the performance of the proposed methods for all images is better than the original of metaheuristic, metaheuristic with chaos, and metaheuristic by Random adjustment, except Lena and Rice images in DE algorithms. In this case, mean objective function of DE1 for images of Lena (0.1577) and Rice (0.2536) are better than the proposed methods, i.e. CDERA1 (Lena: 0.1558, Rice: 0.2529), CDERA2 (Lena: 0.1557, Rice: 0.2533) and CDERA3 (Lena: 0.1573, Rice: 0.2531).

The probabilities of this case, since setting parameters of DE1 are fit with the characteristic of Lena and Rice images. The performance of metaheuristic algorithms depends on their parameter settings. As an example, the best objective function of Lena image (0.1592) is DE1 (F = CR = 0.8). However, the worst objective func-tion of this image (0.1281) is variant of DE1, which is DE5 (F=0.2, CR=0.8). In case of compu-tation time, the best computation time of Lena image is SA1 (43.17 s). However, the worst computation time is the variant of this method, that is SA3 (286.30 s).

Moreover, the performance of metaheuristic algorithms also depends on characteristic of the problem, in this case is images of Lena, Cameraman, Boat and Rice. For example, the best objective function of Lena (0.1590) and Boat (0.1436)

images are CSARA1. However, the best objective function image of Rice (0.2539) is CSARA2 as well as image of Cameraman (0.1703) is CSARA3.

Acknowledgements

This work is supported by Indonesian Directorate General of Higher Education (DIKTI) scholarship BPPDN 2013 for LMRR's study.

References

[1] C. L. J. G. Sauer, M. Rudek, Differential evolution optimization combined with chaotic sequences for image contrast enhancement, Chaos, Solitons and Fractals 42 (2009) 522–529.

[2] C. Munteanu, A. Rosa, Towards automatic image enhancement using genetic algorithms, IEEE Congress on Evolutionary Computation (2000) 1535–1542.

[3] C. Munteanu, A. Rosa, Gray-scale enhancement as an automatic process driven by evolution, IEEE Transaction on systems, man, and cybernetics Part B: Cybernetics 34 (2) 1292–1298.

[4] A. Gorai, A. Ghosh, Gray-level image enhancement by particle swarm optimization, Congress on Nature Biologically Inspired Computing (2009) 72 77.

[5] I. Boussaid, J. Lepagnot, P. Siarry, A survey on optimization metaheuristics, Information Science 237 (2013) 82 117.

[6] El-Ghazali Talbi, Metaheuristics From Design to Implementation, John Wiley Sons, Hoboken, New Jersey, 2009.

[7] J. Liu, J. Lampinen, A fuzzy adaptive differential evolution algorithm, Soft Computing - A fusion of Foundations, Methodologies and Applications 9 (6) (2013) 448–462.

[8] J. Di, N. Wang, Harmony Search algorithm with Chaos for training RBFNN, Academy Publisher, 2013, pp. 2231–2237.

[9] K. Ferens, C. D., K. W., Chaotic simulated annealing for task allocation in multiprocessing system, IEEE International Conference on Cognitive Informatics Cognitive Computing (ICCI-CC), 12 (2012) 26–35.

[10] N. Noman, D. Bollgala, H. Iba, An adaptive differential evolution algorithm, IEEE Evolutionary Computation (2011) 2229–2236.

[11] X. Li, J. Xu, Y. Yang, A chaotic particle swarm optimization - based heuristic for market-oriented task-level scheduling in cloud workflow systems, Computationnal Intellignece and Neuroscience 2015 (2015) 11 pages.

[12] H. Cheng, H. Xu, A novel fuzzy logic approach to contrast enhancement, Pattern Recognition 33 (2000) 809–819.

[13] M. El-Santawy, A. Ahmed, R. El-Dean, Chaotic differential evolution optimization, Computing and Information System Journal 16 (2) (2012) 1–4.

[14] S. Kirkpatrick, C. Gelatt, M. Vecchi, Optimization by simulated annealing, Science, New Series 220 (4598) (1983) 671–680.

[15] B. Suman, P. Kumar, A survey of simulated annealing as a tool for single and multiobjective optimization, Journal of the Operation research Society 577 (2006) 1143–1160.

[16] R. Eberhart, J. Kennedy, A new optimizer using particle swarm theory, Sixth International Symposium on Micro Machine and Human Science (1995) 39–43.

[17] K. S. Lee, Z. W. Geem, A new meta-heuristic algorithm for continuous engineering optimization: harmony search theory and practice, Comput. Methods Appl. Mech. Engrg 194 (2005) 3902–3933.

Jurnal Ilmu Komputer dan Informasi (Journal of Computer Science and Information). 10/2 (2017), 77-85
DOI: http://dx.doi.org/10.21609/jiki.v10i2.418

SENTENCE ORDERING USING CLUSTER CORRELATION AND PROBABILITY IN MULTI-DOCUMENTS SUMMARIZATION

I Gusti A. S. Adi Guna, Suci Nur Fauziah, and Wanvy Arifha Saputra

Informatics Department, Faculty of Information and Technology, Institut Teknologi Sepuluh Nopember, Jl. Raya ITS Kampus Sukolilo, Surabaya, 60111, Indonesia

E-mail: socrates.adiguna@gmail.com, wanvy15@mhs.if.its.ac.id

Abstract

Most of the document summary are arranged extractive by taking important sentences from the document. Extractive based summarization often not consider the connection sentence. A good sentence ordering should aware about rhetorical relations such as cause-effect relation, topical relevancy and chronological sequence which exist between the sentences. Based on this problem, we propose a new method for sentence ordering in multi document summarization using cluster correlation and probability for English documents. Sentences of multi-documents are grouped based on similarity into clusters. Sentence extracted from each cluster to be a summary that will be listed based on cluster correlation and probability. User evaluation showed that the summary result of proposed method easier to understanding than the previous method. The result of ROUGE method also shows increase on sentence arrangement compared to previous method.

Keywords: *Document Summarization, Cluster Ordering, Cluster Correlation, Probability*

Abstrak

Sebagian besar ringkasan dokumen dihasilkan dari metode *extractive*, yaitu mengambil kalimat-kalimat penting dari dokumen. Ringkasan dengan metode *extractive* sering tidak mempertimbangkan hubungan antar kalimat. Pengurutan kalimat yang bagus menunjukkan hubungan *rhetorical,* seperti hubungan sebab akibat, topic yang relevan, dan urutan yang kronologis diantara kalimat. Berdasarkan permasalahan ini, diusulkan sebuah metode baru untuk pengurutan kalimat pada peringkasan dari beberapa documen menggunakan *cluster correlation* dan *probability*untuk dokumen berbahasa inggris. Kalimat dari beberapa dokumen dikelompokkan berdasarkan kemiripannya ke dalam cluster-cluster. Kalimat diekstrak dari setiap cluster untuk menjadi ringkasan, ringkasan akan diurutkan berdasarkan *cluster correlation* dan *probability.* Hasil evaluasi pengguna menunjukkan hasil ringkasan dari metode usulan lebih mudah dipahami dari pada metode sebelumnya. Hasil ROUGE juga menunjukkan peningkatan susunan kalimat dari metode sebelumnya.

Kata Kunci: *Peringkasan Dokumen, Pengurutan Cluster, Cluster Correlation, Probability*

1. Introduction

In the present, huge electronic textual information is available and accessible. The information retrieval technology has make everyone can obtain large number of related documents using search engine. However, this situation also makes people need large of time to obtain the necessary information from all documents that they found. Automatic document summarization has been concern of the researchers for decade to solve this problem [1].

Most of existing automatic document summarization methods are extractive based, which mean they need to find the significant sentence or paragraph in documents, and arrange them to become a summary. However, extractive based summarization has a big hole on it. How the system arranges the sentences (or paragraphs) to become a proper summary is crucial part of this method. Sentence ordering without considering the relation among them can cause in incoherent summary [1].

A good sentence ordering should aware about rhetorical relations such as cause-effect relation, topical relevancy and chronological sequence which exist between the sentences. For example, if sentence A mention the event that caused by sen-

tence B, then we might want to order the sentence A before the sentence B in a summary that contains both sentences A and B [2].

Sentence ordering is more difficult in multiple documents, because the sentences that will draw up the summary is extracted from different writing styles documents and authors, which no one of the document can provide a standard sequence for all sentences that extracted. The way it orders the sentences should be context-aware and represent all of the source documents [3].

The previous research [4-7] proposed summarization document using Similarity Based Histogram Clustering (SHC) and cluster importance for sentence ordering method. SHC method capable to prevent the cluster contain duplicate sentence in it. SHC also prevent the situation that led the summary become redundant. Cluster importance is an ordering cluster method based on frequency information density among cluster's member. The information density is calculated by count the number of terms in cluster that has frequency above the predefined threshold. However, cluster importance ignores structure or relation among the cluster that provides the sentences that form the summary is not associated with each other.

Based on this problem, we propose a new method for sentence ordering in multi document summarization using cluster correlation and probability for English documents. This method is inspired by correlation coefficient method that used to measure the degree of relatedness of two vectors [8].

Sentence Clustering

Sentence clustering is used to find the similarity and dissimilarity across the documents. In sentence clustering, if the number of cluster has been determined, there is possibility that some sentences will be forced to become member of a cluster although it should not be. This probable error in cluster's member placement may cause some clusters to have duplicate member of sentence and led the summary become redundant. To avoid the problem in cluster member placement, we use Similarity Histogram-based Clustering (SHC) for the sentence clustering method. SHC can be used to make clusters by measure the similarity among the sentences [6].

The histogram that used in SHC is statistical representation of the similarity between the cluster members. Each value on histogram shows a certain similarity interval. A Similarity threshold is used in cluster member's selection. Every sentence will be registered into an appropriate cluster. The appropriate cluster can be found only if the sentence does not make the similarity value of these cluster members reduce. If that sentence can't be fit on any cluster, it should make a new cluster by itself [4]. The uni-gram matching similarity is used as function of similarity of two sentences, S_j and S_i. The similarity between sentences is calculated by count the corresponding words between S_j and S_i. ($|s_i| \cap |s_j|$). Then it is divided by the total length of the words that form S_j and S_i ($|s_i|+|s_j|$) as shown in equation(1).

$$sim(s_i, s_j) = \frac{(2 * | s_i | \cap | s_j |)}{| s_i | + | s_j |} \quad (1)$$

$Sim = \{sim_1, sim_2, sim_3, \ldots. sim_m\}$ is collection of similarity between a couple sentences with $m= n(n=1)/2$. The equation determines the histogram function equation(2),

$$h_i = count(sim_j) \quad \text{for } sim_{li} \leq sim_j \leq sim_{ui} \quad (2)$$

Where sim_{li} shows the minimum similarity bin to-i and sim_{ui} is maximum similarity bin to-i. Histogram Ratio (*HR*) of cluster calculated by equation (3) and threshold determined by equation (4).

$$HR = \frac{\sum_{i=T}^{n_h} hi}{\sum_{j=1}^{n_b} hj} \quad (3)$$

$$T = [S_T * n_b] \quad (4)$$

S_T is similarity *threshold*, where bin number that corresponds to the similarity threshold (S_T) annotated with *T*.

Cluster Correlation

The inter-cluster correlation calculation based on frequencies a term that contained in each cluster. Some words are not always available in each cluster, that make so many comparisons with zero that does not affect to the results. For that, we simplify the calculation by use only important words to calculate the correlation of the cluster.

The important words are the words that often appear in all document. Determination of important words based on the frequency of occurrence terms that meet the threshold (θ) in all documents. The inter-cluster correlation is calculated by determine the cluster *a* and $T = \{t_1, t_2, t_3, \ldots t_n\}$ is set of term in cluster *a*. $w_{x,ti}$ is number of frequency term $t_i \in T$ in cluster *a,* then *a* has member $a = \{w_{a,t_1}, w_{a,t_2}, \ldots. w_{a,t_n}\}$. Weight of term frequency

in cluster *a* and *b* ($w_{t\ ab}$) calculated by multiply the number of distinct term *n* with the total of the result of multiplication the weight of each term t_i in cluster *a* (w_{a,t_i}) with the weight of each term t_i in cluster b (w_{b,t_i}), as the equation(5).

$$w_{t\ ab} = n \sum_{i=1}^{n} w_{a,t_i} * w_{b,t_i} \quad (5)$$

The Term frequency in cluster *a* and b calculated by multiply the cluster term frequency TF_a and TF_b in equation (6),

$$TF_a = \sum_{i=1}^{n} w_{a,t_i} \quad (6)$$

$$TF_b = \sum_{i=1}^{n} w_{b,t_i} \quad (7)$$

$$TF_{a,b} = TF_a * TF_b, \quad (8)$$

Correlation coefficient is generated by reduce the weight term frequency $w_{t\ ab}$ (equation (5)) with the term frequency of all cluster $TF_{a,b}$ (equation(8)), then it was divided by square root of weight term of every cluster. The function to calculate correlation is shown in the equation(9):

$$r_{(\vec{ta},\vec{tb})} = \frac{w_{t\ ab} - TF_{a,b}}{\sqrt{\left[n \sum_{i=1}^{n} w_{a,t_i}^2 - TF_a^2\right]\left[n \sum_{i=1}^{n} w_{b,t_i}^2 - TF_b^2\right]}} \quad (9)$$

The value of correlation has ranges from -1 (strong negative correlation) to 1 (strong positive correlation). If the value of correlation is 0, it is mean it has not any correlation with the other cluster [8]. The correlation coefficient was used as the basis for calculate cluster order on this paper.

Distributed Local Sentence

Distributed local sentence is a method of sentence arrangement. This sentences distribution method was proposed method by [4]. Distributed local sentence method is constructed through calculated the probability distribution, calculated sum of distribution, calculated expansion of distribution, calculated the weight of the component sentences and calculated the weight distributed local sentence. The weight local sentence is obtained by summing the entire component forming sentence *i* in cluster *k* which is divided by the number of component forming sentence *i* in cluster *k*. The equation (10) to calculate weight of distributed local sentence.

$$W_{ls}(s_{ik}) = \frac{1}{|S_{ik}|} \sum_{W_{t1,jk} \in S_{ik}} Wt_{l,jk} \quad (10)$$

2. Methods

There are five main steps that used in this research, i.e. preprocessing, sentence clustering, sentence ordering, sentence extraction and sentence arrangement. Figure 1 shows the process to generate a summary. The research method was adopted from the research [4].

Preprocessing

This process aims to prepare the data which is used in the sentence clustering. The preprocessing process consist of tokenizing, stopword removal, and stemming. Tokenizing is process of beheading the sentence into standalone words. Stopword removal is process of removing unused words. Stemming is process of getting the words to lower-case form. In this research, tokenizing process uses Stanford Natural Language Processing, stopword removal process uses stoplist dictionary and stemming process uses Library English Porter Stemmer [4].

Sentence Cluster

Data which is generated from preprocessing process will be grouped by using Similarity-based Histogram Clustering (SHC). SHC method was chosen because this method ensures the results of cluster remain convergent. Similarity between

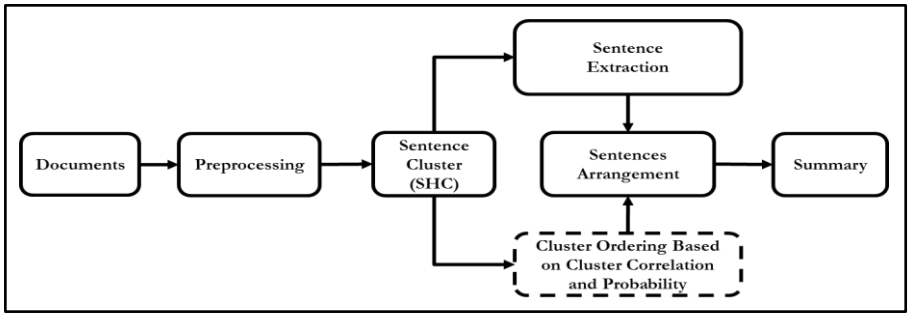


Figure 1. Sentence ordering based on cluster correlation

sentences was calculated by using uni-gram matching-based similarity as equation(1-4).

A sentence can be entered on certain cluster if it passes the criteria of the cluster. But if the sentence does not meet the criteria of all existing cluster, it will set up a new cluster. The SHC method is used in this research was adapted from research [4].

Cluster Ordering

The proposed method or contribution of this research is on the cluster ordering section. Sorting is useful to determine the order of sentences that represent in the summary. The summary will be easily understood if the sentences sorted according to the proximity of it with the topic or content.

Cluster ordering is used to calculate the order of each cluster using correlation between cluster and probability each cluster. There are three steps in this section, i.e. inter-cluster correlation calculation, probability calculation and weighted coherence.

Cluster Correlation

After sentence clustered using SHC, the first step was to determine the correlation of sentence cluster. The cluster correlation formula that adopted was equation (9) from correlation coefficient method [8] which can be used to measure the degree of relatedness for two vectors. That correlation can assume the value of correlation has ranges from -1 (strong negative correlation) to 1 (strong positive correlation). If the value of correlation is 0, it means that it uncorrelated with the other cluster.

In this research, the correlation was used without regard it has positive or negative value. So, in this research we add absolute value as the equation (11).

$$Correlation_{(\vec{ta},\vec{tb})} = |r_{(\vec{ta},\vec{tb})}| \qquad (11)$$

Correlation cluster values between the two clusters will determine the order of the cluster. In equation (11), the value of correlation has ranges from 0 to 1. If the value correlation is 0, it means has not any correlation with other cluster, but if the value reach 1, it means has any correlation regardless positive or negative correlation as equation (9).

Cluster Probability

The second step, to determine the probability. Probability calculated by use the frequency of occurrence of important sentences. That means Importance sentence is a sentence which consists of number of important words on it. Some cluster with large of member should have more weight because they have sentence information density.

It is difference with the cluster that has one or little members. Therefore, it required probability calculation. The Probability value obtained by comparing the frequency of important sentences in *cluster a* with frequency of importance sentence in all document (12).

$$\mathrm{P}(cluster_x) = \frac{freqImpSentece_{cluster\,x}}{freqImpSentece_{all\,document}} \qquad (12)$$

Weight Coherency

The third step, Weight coherency will determine the order of the sentence, the highest coherence weight will be concept of sentence ordering. WC_a is weight coherency which obtained by multiplying the sum of correlation value *a* to *x* with probability cluster *a,* where cluster *x* is correlation pair of cluster *a,* as equation(13).

$$WC_a = \sum_{x=1}^{n} Correlation_{a,x} \times P(a) \qquad (13)$$

The equation(13) can be assumed correlation cluster in equation(11) multiply by cluster probability in equation(12). The weight coherency can stabilize the value of correlation and frequency importance word in sentence within cluster. So the cluster that have most important sentences will be select as priority and the other cluster follow with sorted by descending based on weight coherency value.

It different in equation(9) that just only calculate cluster correlation and not focus the importance word in sentence within cluster. So the result of cluster ordering just sorted by high value of correlation in descending. The effect of that, negative correlation become last priority, and althought it have negative correlation, it still have many importance sentence rather than zero correlation.

Sentence Extraction

Sentence extraction is a phase to select a sentence that represents the cluster. The extracted sentences then used to arrange the summary. Sentence extraction uses sentence distribution from researcher [4]. Sentence distribution is used to determine the position of each sentence in the cluster, if a sentence that has elements spreader in a cluster will have a higher position in the cluster. The equation-

$$ROUGE - N = \frac{\sum_{S \in Summ_{ref}} \sum_{N-gram \in S} Count_{match}\,(N-gram)}{\sum_{S \in Summ_{ref}} \sum_{N-gram \in S} Count\,(N-gram)} \quad (14)$$

(10) is used for sentence distribution that used in this research. Sentence distribution that has the highest weight on each cluster will be used to arrange the summary.

Sentences Arrangement

The extracted sentence that used to arrange the summary will be listed based on the cluster ordering result. Every cluster contains one or more sentence that chosen based on sentence extraction section using sentence distribution method. Number of cluster is equal to number of sentence in summary

Evaluation method

The evaluation of automatic summarization is use ROUGE (Recall Oriented Understudy for Gisting Evaluation) method. ROUGE-N measures the ratio of n-grams between the candidate summary and the set of reference summaries. ROUGE is effective to evaluate the document summary result.

ROUGE-N is computed as equation (14) [9]. N is the length of N-gram, $Count_{match}$ (N-gram) is the maximum number of N-gram between the candidate summary and the set of reference summaries. Count(N-gram) is the number of N-gram in reference summaries. In this research, we use ROUGE-1 and ROUGE-2 where the best condition of ROUGE-1 and ROUGE-2 is 1.

3. Result and Analysis

Data Set

In this research, we use DUC (document understanding conferences) 2004 task 2 dataset from http://www-nlpir.nist.gov/projects/duc/data/2004-_data.html. DUC is one of the most popular data set for document summary. It is consisting of news documents collection from Associated Press and New York Times. 25 topics dataset is used where every topic consists of 10 documents.

Results

The experiment is performed using java programming. We compare the proposed method result with cluster importance method from [4]. The research [4] uses sentence clustering (SHC) and sentence distribution method for summary extraction and uses cluster importance for ordering sentence.

There are four parameters in this research, i.e. HR_{min}, ε, S_T, and α. HR_{min}, ε, and S_T are parameter for sentence clustering with SHC method. And then parameter α is parameter for sentence extraction. This research uses HR_{min}= 0.7 ,ε = 0.3, S_T = 0.4, θ = 10 , and α=0.4 or α=0.2. The other comparator method is also using parameter θ=10 for cluster ordering with cluster importance method.

Our focus in this research is to make a proper summary which more readable. We spread form survey to 20 volunteers and use ROUGE method. The form survey consists 25 topics of generated summary, all topics has been evaluated by 20 volunteers (post graduate students and English teacher) for evaluation. For each topic, every volunteer chooses the most coherent among the summaries which generated using cluster importance method, cluster correlation method, and cluster correlation + probability method.

Figure 2 shows the generated summary by cluster correlation method and cluster importance method for topic "Lebanese presidential election". 12 volunteers out of 20 volunteers have chosen summary using cluster correlation method with probability and another 8 volunteers have chosen cluster correlation method without probability for this topic.

The resulting summary using cluster correlation without probability have similar sentence ordering with the resulting summary using cluster correlation with probability. But, both have very different sentence ordering with the resulting summary using cluster importance.

In our analysis, there are some reasons that make the resulting summary using cluster correlation is better than the resulting summary using cluster importance, as follows: The cluster correlation can make chronological sequence of summary which exist between the sentences, The words in previous sentence are described in the next sentence.

Cluster correlation can avoid pronouns usage in the first sentence; because of the pronouns usage in the first sentence make the summary difficult to understand.

Table 1 shows the number of volunteers which choose cluster importance or cluster correlation as the satisfactory method for each topic. Figure 3 shows the volunteer's choice for cluster importance, cluster correlation with probability and cluster correlation without probability. Table 1 and figure 3 show that most of volunteers chose

Summarization's result using cluster correlation without probability

[1] Parliament on Thursday formally elected Gen. Emile Lahoud, the popular army commander who has the backing of powerful neighbor Syria, as Lebanon's next president.
[2] Lahoud's nomination complies with a tradition that the president be a Maronite Christian, the prime minister a Sunni Muslim and Parliament speaker a Shiite Muslim.
[3] Prime Minister Rafik Hariri, the business tycoon who launched Lebanon's multibillion dollar reconstruction from the devastation of civil war, said Monday he was bowing out as premier following a dispute with the new president.
[4] Such political disputes in Lebanon in the past were solved only with the intervention of Syria, the main power broker in this country.
[5] The new president will be sworn in Nov. 24, the day Hrawi leaves office.
[6] ``Congratulations, your excellency the general," Lebanese Prime Minister Rafik Hariri told army commander Emile Lahoud in a telephone conversation Monday that was headlined on the front-page of the leftist newspaper As-Safir.
[7] But many legislators, who in the past gave their overwhelming support to Hariri, did not name him and, instead, left it to the president to select a prime minister.
[8] A Cabinet minister and a close Syria ally on Wednesday criticized the Syrian-backed choice of the army commander as president, and said he will boycott a vote to elect the military man for the executive post.
[9] Lahoud pledged in a tough inauguration speech to clean up the graft-riddled administration.
[10] Lahoud, a 62-year-old naval officer, enjoys wide public and political support at home and has good relations with Syria.

Summarization's result using cluster correlation with probability

[1] Parlement on Thursday formally elected Gen Emile Lahoud, the popular army commander who has the backing of powerful neighbor Syria, as Lebanon's next president.
[2] Lahoud's nomination complies with a tradition that the president be a Maronite Christian, the prime minister a Sunni Muslim and Parliament speaker a Shiite Muslim.
[3] Prime Minister Rafik Hariri, the business tycoon who launched Lebanon's multibillion dollar reconstruction from the devastation of civil war, said Monday he was bowing out as premier following a dispute with the new president.
[4] Such political disputes in Lebanon in the past were solved only with the intervention of Syria, the main power broker in this country.
[5] The new president will be sworn in Nov. 24, the day Hrawi leaves office.
[6] But many legislators, who in the past gave their overwhelming support to Hariri, did not name him and, instead, left it to the president to select a prime minister.
[7] ``Congratulations, your excellency the general," Lebanese Prime Minister Rafik Hariri told army commander Emile Lahoud in a telephone conversation Monday that was headlined on the front-page of the leftist newspaper As-Safir.
[8] A Cabinet minister and a close Syria ally on Wednesday criticized the Syrian-backed choice of the army commander as president, and said he will boycott a vote to elect the military man for the executive post.
[9] Lahoud pledged in a tough inauguration speech to clean up the graft-riddled administration.
[10] Lahoud, a 62-year-old naval officer, enjoys wide public and political support at home and has good relations with Syria.

Summarization's result using cluster Importance

[1] His word is referred to in the Beirut media as "the password".
[2] Criticism of the nomination process also came from a meeting of Catholic bishops on Wednesday.
[3] Eleven deputies were absent.
[4] All the 118 legislators present at the session cast votes in his favor.
[5] The leading An-Nahar and other newspapers said the delay could last for days.
[6] The two leaders met Friday, but no presidential decree followed.
[7] It added that the money markets remained stable.
[8] But the dispute between the two leaders appears to be over who will have the upper hand in governing the nation of 3.2 milllion.
[9] '"I'm not a candidate, " Harari said in a live interview with CNN.
[10] Lahoud's nomination ends weeks of suspense over the identity of the next head of state.

Figure 2. Topic: "Lebanese Presidential Election"

cluster correlation with probability and cluster correlation without probability as the satisfactory system.

In this research, we used ROUGE-1 and ROUGE-2 as the metric evaluate the difference between ordering generated by automatic summarization and human. The Average ROUGE's score of the proposed method (cluster correlation and probability) and cluster important are presented in Table 2. It shows the comparison of ROUGE's score between the proposed method (cluster correlation and probability) and cluster importance where the proposed method gets better score than cluster importance.

The ROUGE-1's score comparison between cluster importance and proposed method (cluster correlation and probability) of each document are presented on Figure 4. Average documents show adjacent values between cluster importance and proposed method (cluster correlation and probabi-

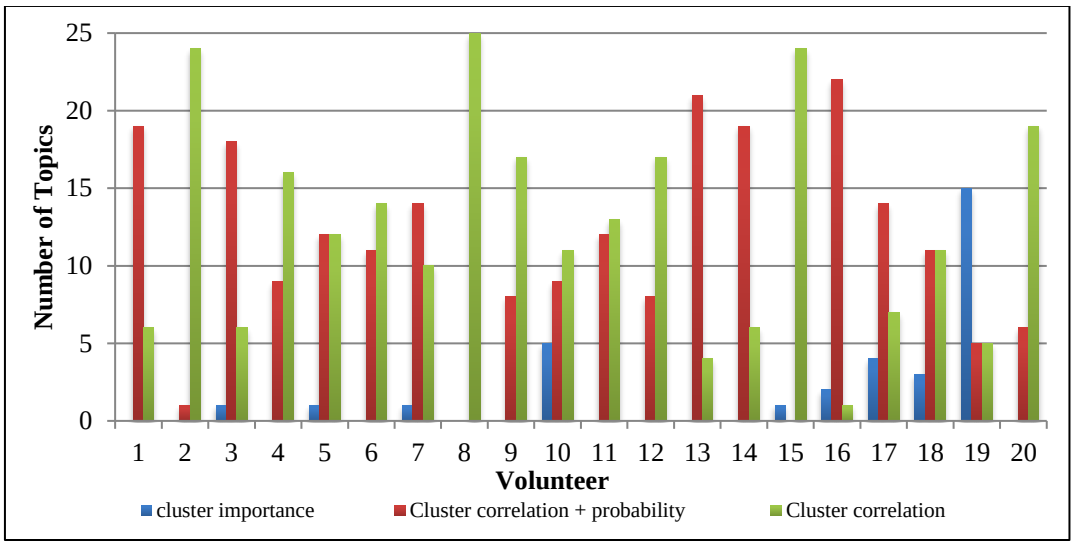


Figure 3. Graph showing the volunteers choice

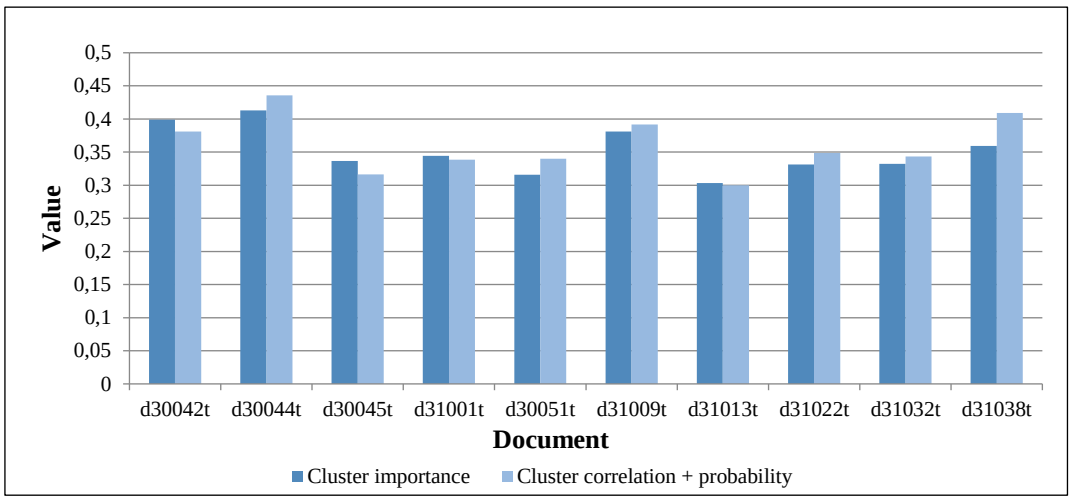


Figure 4. Graph Showing ROUGE 1 Every Document

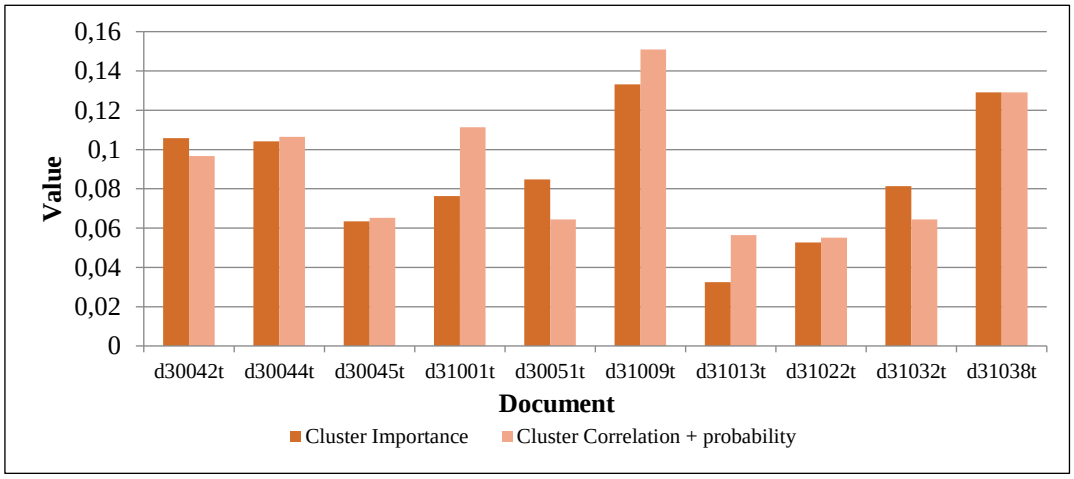


Figure 5. Graph showing ROUGE 2 every document

-lity). The comparison of ROUGE-2 score is presented in Figure 5. Figure 5 shows that proposed method gets better value than cluster importance in some documents.

TABLE 1
SURVEY RESULT FOR EACH TOPIC

Topic	Cluster importance	Cluster correlation + probability	Cluster correlation
T1	2	10	8
T2	1	8	11
T3	1	8	11
T4	3	7	10
T5	1	11	8
T6	0	11	9
T7	0	6	14
T8	2	7	11
T9	0	11	9
T10	1	7	12
T11	2	9	9
T12	1	7	12
T13	2	8	10
T14	1	11	8
T15	2	9	9
T16	3	11	6
T17	3	8	9
T18	2	8	10
T19	2	9	9
T20	1	11	8
T21	1	11	8
T22	2	11	7
T23	1	5	14
T24	0	12	8
T25	2	8	10

TABLE 2
TESTING OF SENTENCE ORDERING

Sentence Ordering Method	ROUGE 1	ROUGE 2
Cluster correlation + probability	0.360	0.090
Cluster importance	0.351	0.086

4. Conclusion

The proposed method is sentence ordering using cluster correlation and probability in multi document summarization has been successful. The result showed that the proposed method gives better summary than cluster importance. Summary using cluster correlation is preferred by most of volunteers.

Cluster correlation method has proven better than cluster importance which has average score 0.360 for ROUGE 1 and 0.090 for ROUGE 2 in Table 2. There is an increase value by 0.004 on ROUGE 1 and ROUGE 2. The increase in numbers is due to the evaluation on the ground-truth that measured from the summary of documents based on important sentences regardless of its sentence ordering, to know the sentences ordering in the document objectively and more significant, then the evaluation that used was human perceptions like surveys.

From the ordering of sentences, the cluster importance method not consider the correlation between important sentences in the cluster. Then the result of sentences ordering is different as shown in Figure 2, Figure 3 shows the proposed method is more objective on a topic than cluster importance method.

Combination cluster correlation with probability generate similar summary with cluster correlation without probability. The usage of probability in this method has no effect on the first to fifth sentence of the summary results. The resulting summary from the cluster correlation method represents not only the topic of document, but also the chronological sequence which exists among the sentences.

In the sentence ordering of multi document summarization, source documents cannot provide enough information in sentence ordering of summary as ground truth for evaluation. Figure 1 explains sentence ordering method of this research. The proposed method consists four steps:

The first step, preprocessing for prepare data which used sentence clustering. The second step, document is categorized using SHC method and produces clusters. The third step, sentence extraction uses sentence distribution method. The fourth, sentence ordering uses cluster correlation between clusters.

Because of sentence ordering is non-standard of ordering method; it is difficult to make a proper order of the sentence. The summary is extracted from different writing styles documents and authors. Our method is formed based on the information of source documents, which must be in multi document summarization.

The results show that the proposed method can improve the quality of the summary from multi documents which use SHC as sentence clustering method. It is related with some of previous researches which used SHC clustering method [4]–[7].

In future work, we will focus in how to deal with a large amount of document. The large quantity of document can make a huge amount of cluster, which makes the complexity of correlation calculation will be increased and hard to handle.

References

[1] U. Hahn and I. Mani, "Challenges of automatic summarization," *Computer (Long. Beach. Calif).*, vol. 33, no. 11, pp. 29–36, 2000.

[2] D. Bollegala, N. Okazaki, and M. Ishizuka, "A bottom-up approach to sentence ordering for multi-document summarization," *Inf. Sci. (Ny).*, vol. 217, no. 1, pp. 78–95, 2012.

[3] G. Peng, Y. He, N. Xiong, S. Lee, and S. Rho, "A context-aware study for sentence ordering," *Telecommun. Syst.*, vol. 52, no. 2, pp. 1343–1351, 2013.

[4] Wahib, A., Arifin, A.Z. and Purwitasari, D., 2016. "Improving Multi-Document Summary Method Based on Sentence Distribution". *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 14, no.1, pp.286-293.

[5] Suputra H. G.I, Arifin Z.A, and Y. A, "Stategi Pemilihan Kalimat pada Peringkasan Multi-Dokumen Berdasarkan Metode Clustering Kalimat," *J. Ilmu Komput.*, 2013.

[6] K. Sarkar, "Sentence Clustering-based Summarization of Multiple Text Documents," *Tech. – Int. J. Comput. Sci. Commun. Technol.*, vol. 2, no. 1, pp. 325–335, 2009.

[7] R. Azhar, M. Machmud, H. A. Hartanto, and A. Z. Arifin, "Pembobotan Kata Berdasarkan Klaster pada Optimisasi Coverage , Diversity dan Coherence untuk Peringkasan Multi Dokumen."

[8] S. Al-anazi, H. Almahmoud, and I. Al-turaiki, "Finding Similar Documents Using Different Clustering Techniques," in *Procedia - Procedia Computer Science*, 2016, vol. 82, no. March, pp. 28–34.

[9] C. Y. Lin, "Rouge: A package for automatic evaluation of summaries," *Proc. Work. text Summ. branches out (WAS 2004)*, no. 1, pp. 25–26, 2004.

Jurnal Ilmu Komputer dan Informasi (Journal of a Science and Information). 10/2 (2017), 86-95
DOI: http://dx.doi.org/10.21609/jiki.v10i2.449

PSNR BASED OPTIMIZATION APPLIED TO ALGEBRAIC RECONSTRUCTION TECHNIQUE FOR IMAGE RECONSTRUCTION ON A MULTI-CORE SYSTEM

A. Bharathi Lakshmi[1] and D. Christopher Durairaj[2]

[1]Department of Information Technology, Assistant Professor, Affiliated to Madurai Kamaraj University, Virudhunagar, 626001, India.
[2]Research center in Computer Science, VHNSNC, Virudhunagar, 626001, India.

E-mail: bharathilakshmi.a@gmail.com, kesterkaren@gmail.com

Abstract

The present work attempts to reveal a parallel Algebraic Reconstruction Technique (pART) to reduce the computational speed of reconstructing artifact-free images from projections. ART is an iterative algorithm well known to reconstruct artifact-free images with limited number of projections. In this work, a novel idea has been focused on to optimize the number of iterations mandatory based on Peak to Signal Noise Ratio (PSNR) to reconstruct an image. However, it suffers of worst computation speed. Hence, an attempt is made to reduce the computation time by running iterative algorithm on a multi-core parallel environment. The execution times are computed for both serial and parallel implementations of ART using different projection data, and, tabulated for comparison. The experimental results demonstrate that the parallel computing environment provides a source of high computational power leading to obtain reconstructed image instantaneously.

Keywords: *Image Processing, Image Reconstruction, Iterative Image Reconstruction, Algebraic Reconstruction Technique, Parallel Processing, OpenMP*

Abstrak

Pekerjaan saat ini mencoba untuk mengungkapkan Teknik Rekonstruksi Algebraic paralel (pART) untuk mengurangi kecepatan komputasi untuk merekonstruksi gambar bebas artifak dari proyeksi. ART adalah algoritma iteratif yang dikenal untuk merekonstruksi gambar bebas artefak dengan jumlah proyeksi yang terbatas. Dalam karya ini, sebuah gagasan baru difokuskan untuk mengoptimalkan jumlah iterasi yang wajib berdasarkan Peak to Signal Noise Ratio (PSNR) untuk merekonstruksi gambar. Namun, ia menderita kecepatan perhitungan terburuk. Oleh karena itu, upaya dilakukan untuk mengurangi waktu komputasi dengan menjalankan algoritma iteratif pada lingkungan paralel multi-core. Waktu eksekusi dihitung untuk penerapan ART secara serial dan paralel dengan menggunakan data proyeksi yang berbeda, dan, ditabulasikan sebagai perbandingan. Hasil percobaan menunjukkan bahwa lingkungan komputasi paralel menyediakan sumber daya komputasi tinggi yang menghasilkan gambar yang direkonstruksi seketika.

Kata Kunci: *Pemrosesan gambar, rekonstruksi gambar, rekonstruksi gambar iteratif, teknik rekonstruksi aljabar, pemrosesan paralel, OpenMP*

1. Introduction

Image reconstruction methods are central to many of the new applications of medical imaging such as Positron Emission Tomography (PET), Computed Tomography (CT), Magnetic Resonan-ce Imaging (MRI) and Electron Magnetic Reso-nance Imaging (EMRI). They are most commonly used to visualize detailed internal structure and limited function of the object of interest.

Image reconstruction is a mathematical process that generates images from projection data acquired at many different angles around the object of interest. The projections are collected by sweeping the magnetic field at projection angles defined by the magnetic field gradient directions [1, 2]. To perform image reconstruction, the projections $p_\theta(r)$, collected along a set of field-gradient orientations in polar coordinates, are used to obtain the image *f(x, y)* [3] as given in the equation(1).

$$\begin{aligned} & f(x,y) \\ & = \int_0^\pi P_\theta^*(r)d\theta \\ & = \int_0^\pi [\int_{-\infty}^{\infty} P_\theta(k)|k|e^{-2\pi ikr} - dk]d\theta \end{aligned} \quad (1)$$

Here r is taken on the x-y plane such that $r = xcos\theta + ysin\theta$, and $p_\theta^*(r)$ is the projection $p_\theta(r)$

filter according to the expression inside the square brackets [3].

Image reconstruction has been carried out using different types of reconstruction algorithms [4, 1]. Reconstruction methods utilize projection data as input and generate the estimate that resembles the internal structure as output [5, 6]. Data sets with 36 projections measured from 0^0 to 180^0 around the phantom object were considered in the present study. The same data set was used for testing the capability of the algorithms from restricted number of projections, by skipping projections at uniform angular distribution.

Reconstruction of images is usually done in two ways: Analytical and Iterative. Analytical method such as Back Projection (BP) or Filtered Back Projection (FBP) is used for different imaging modalities such as CT and PET in clinical settings because of its speed and easy implementation [3]. For noisy projection data as well as for limited number of projections, the FBP method of image reconstruction shows very poor performance. Hence currently there is considerable interest to evaluate the use of other reconstruction methods for medical imaging techniques [6]. FBP algorithm produces high-quality images with excellent computational efficiency. However, FBP produces low Signal-to-Noise Ratio (SNR) images when limited number of projections is used [12].

An Iterative method using a non-linear fit to the projection data has shown to give ripple free images [7]. Iterative Methods are based on optimization strategies incorporating specific constraints about the object and the reconstruction process. The iterative reconstruction techniques perform better than the FBP method when reconstruction is attempted with limited number of projection data [3]. Some of the accepted iterative algorithms are Additive Algebraic Reconstruction Technique (AART) and Multiplicative Algebraic Techniques (MART) [12].

However, the quality of the reconstructed images obtained from AART algorithm depends on number of iterations. Based on the number of available number of projections and the size of the phantom, the number of iterations differs. It is therefore necessary to find the best iteration in order to exploit correctly the promising iteration based on a better Peak to Signal Noise Ratio (PSNR) of the reconstructed images. Based on the equation(1) an optimization program has been developed for the given data set. The best PSNR value is obtained and verified whether the same PSNR value is achieved even after the selected iteration.

Parallel computing is emerging as a principle theory in high performance computing [14]. In recent years, parallel computing with massive data has emerged as a key technology in imaging techniques also. Shared memory parallelization has been proved to be a best way to attain better runtime performance recently for image reconstruction [15]. A shared-memory multiprocessor (SMP) consists of a number of processors accessing one or more shared memory modules. The penalty of using inter-processor communication is not up to the mark on SMP compared to distributed memory architectures [15]. For a relatively large data size, it is advantageous to use SMP architecture. It has also been shown that shared memory parallelization is more suitable than distributed memory parallelization for image processing tasks and leads to better throughput as most of the computers now have two or more processors which share the memory [16]. These features have motivated us to perform the parallelization of Algebraic Reconstruction Technique (ART) on a SMP parallel architecture.

The present study focuses on reducing the computational complexity of ART using parallel programming techniques. Section 2 describes about ART briefly. The design and implementation ART algorithm in both parallel and sequential version are given in section 3 Section 4 discusses the results.

2. Methods

Radon Transformation

The main application of image reconstruction from projection technique is mostly related to medical image processing. The Procedure to implement Image Reconstruction from Projection (IRP) technique in the practical applications are "scanning" or "data acquisition" is considered to be the first and the very important step [17]. Such data acquisition is done by means of PET, CT, MRI or EMRI in a procedure by passing rays in specific intervals of angles.

The Radon Transformation is a fundamental tool that computes projections of an image matrix along specified directions [18]. The 2D Radon transformation is the projection of the image density along a radial line oriented at a specific angle. The value of a 2-D function at an arbitrary point is uniquely obtained by the integrals along the lines of all directions passing the point. The Radon transformation shows the relationship between the 2-D object and its projections. Figure 1 shows a 2-D function $f(x, y)$. Integrating along the line, whose normal vector is in θ direction on s axis results in the $g(s, \theta)$ projection represented in

equation(2). The points on the line whose normal vector is in θ direction and passes the origin of (x, y)-coordinate satisfy the equa-tion $xcos\theta + ysin\theta = 0$. The general equation of the radon transform is acquired as

$$g(s,\theta) = \iint f(x,y).\delta(xcos\theta + ysin\theta - s)dxdy \quad (2)$$

where δ is zero for every argument except to 0 and its integral is one [19].

The projection data obtained thus from Radon Transformation is utilized as input by the Reconstruction algorithm that produce estimates of the original internal structure as output [5, 20]. The size of data sets acquired by the different imaging modalities are usually huge because of the complex data type of the raw collection data, multiple gradients in the experiments, high dimen-sions of the resultant 3-D images, higher k-space requirement of whole body imaging and the number of points collected from the imager. The iterative methods, hence, suffers more recon-struction time.

Algebraic Reconstruction Technique (ART)

Image reconstructions based on Iterative methods create two-dimensional images from scattered or incomplete projections such as the radiation readings acquired during a medical imaging study. Algebraic Reconstruction Technique (ART) falls under the category of Iterative methods.

ART is one of the methods used for solving the linear system which appears in image reconstruction. ART can be broadly classified as either sequential or simultaneous or block itera-tive [21]. ART is a fully sequential method and has a long history and literature. Originally it was proposed by Kaczmarz [22], and independently, for use in image reconstruction by Gordan, Bender and Herman [23]. The vector of unknowns is updated at each equation of the system, after which the next equation is addressed. If system of equation is (0.1) consistent, ART converges to a solution of this system. If the system is incon-sistent, every sub-sequence of cycles through the system converges, but not necessarily to a least square solution [24].

ART perform corrections during iterations, without increasing the computation time. The image $\boldsymbol{f(x,y)}$ is a continuous two dimensional function and an infinite number of projections are mandatory for reconstruction [12]. In practice $\boldsymbol{f(x,y)}$ is calculated using a finite number of points f_j $(j = 1, 2, 3, \dots, N)$ where N represents the total number of cells, from a finite number of projections as shown in figure. 2.

In figure 2 a ray is a fat line running through the (x, y)-plane where each ray is of width *r*. A line integral is called a ray-sum represented as p_i measured with ith ray as shown in figure. 2.

The relationship between the $f_j's$ and $p_i's$ may be expressed as

$$\sum_{j=1}^{N} w_{ij} f_j = p_i, \qquad i = 1, 2, \dots, M \quad (3)$$

where M is the total number of rays (in all the projections) and w_{ij} is the weighting factor that represents the contribution of the jth cell to the ith ray integral and p_j represents a set of matrix equation for the data point f_j.

The expanded form for equation(3) for the jth sample is given by

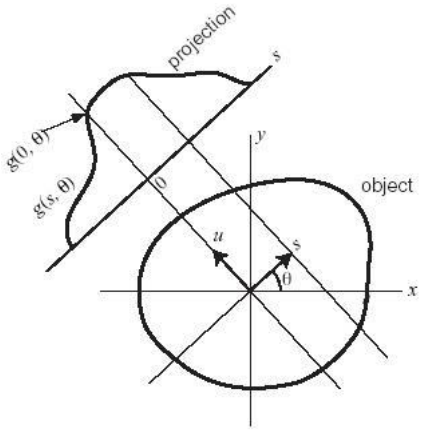


Figure. 1. The Radon Transform computation

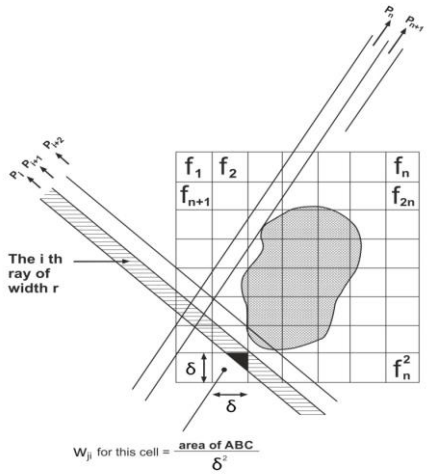


Figure. 2. Representation of an image projected on ith ray.

$$\begin{aligned} w_{11}f_1 + w_{12}f_2 + w_{13}f_3 + \ldots + w_{1N}f_N &= p_1 \\ w_{21}f_1 + w_{22}f_2 + w_{23}f_3 + \ldots + w_{2N}f_N &= p_2 \\ &\vdots \\ w_{M1}f_1 + w_{M2}f_2 + w_{M3}f_3 + \ldots + w_{MN}f_N &= p_M \end{aligned} \quad (4)$$

Equation(4) can also be expressed in the form of algebraic equations as

$$\begin{aligned} P_j &= W_{1j}f_1 + W_{2j}f_2 + W_{3j}f_3 + \ldots + W_{nj}f_n \\ P_j &= \sum_{j=1}^{N} W_{ij}f_j \qquad i = 1, 2, \ldots, M \end{aligned} \quad (5)$$

Here, W_{ij} is the weighting factor that represents the contribution of the j^{th} cell to the j^{th} sample sum and P_j represents a set of matrix equations for the data point f_j. Most of the w_{ij} in Eqn. 4 is zero since only a small number of cells contribute to any given ray-sum. The density values f_j are iteratively adjusted until the calculated projections agree with the measured projections [12]. Each projected density is thrown back across the reconstruction space in which the densities are iteratively modified to bring each reconstructed projection into concur with the measured projection [25]. The projection data set is sustained in a vector and a weight sparse matrix w_{ij} is constructed. Every row in w_{ij} sparse matrix may contain $m + n - 1$ (where m x n is the resolution). As every row stands for the length of the segments obtained by the intersection of ray with the grid, and all reconstruction algorithms use rows of sparse matrix, the best method to store this matrix is in compressed row storage [17].

For each sample, the correction coefficient is computed as:$\alpha_i = \sum_{j=1}^{N} W_{ij}^2$. The average value of the correction coefficient is calculated. Correction is applied for each cell j as given:$f_i^{l-1} + \lambda\overline{\Delta P_j}$, where λ is the relaxation parameter. This procedure is iteratively performed for all of the projection angles. As the size of the data set increases, the computation time increases.

OpenMP Architecture and Directives

Parallel computing is a form of computation in which many calculations are carried out simultaneously; large problems are divided into smaller ones, solved concurrently. The parallelism can be applied in image processing applications by three main ways: 1) Data Parallel 2) Task Parallel and 3) Pipeline Parallel. In Data Parallel approach, the data is divided and distributed among the computing units. The data parallelism to image data can be applied using one of three basic ways: i) Pixel Parallel ii) Row or Column parallel and iii) Block Parallel [27]. This algorithm is parallelized in row/column parallel. In task parallel, image processing instructions/low level operations are grouped into tasks and each task is assigned to a different computational unit. If image processing application requires multiple images to be processed, then pipeline processing of images can be done [28].

pART is implemented using OpenMP parallel computing in C language. OpenMP is a programing model for SMP computer systems. Data in memory can either be shared between all threads or private for one thread. Data transfer between threads is transparent to the programmer. OpenMP uses fork-and-join model of parallel execution. The program written with OpenMP begins execution as a single-process, called the master thread. The master thread executes the current program sequentially until it bump into parallel directives such as #pragma omp. The master thread forks a number of worker threads when it enters a parallel region. A parallel region is a block of code that is executed by all threads concurrently.

The “parallel for” or “for” is a wok sharing directive that distributes the workload of a “for” loop among all the threads. Data sharing of variables is mentioned at the beginning of the parallel region or work sharing construct using the SHARED or PRIVATE Clauses.

Data Set

The reconstruction system uses Shepp Logan phantom data of different sizes such as 64, 128 and 256. The figure. 3(a), figure. 3(b), figure. 3(c) shows the Shepp Logan phantom image of 64x64, 128x128 and 256x256 sizes respectively.

The projection data of the phantom images

Figure. 3. The Shepp Logan Phantom Image of size (a)64x64 (b)128x128 (c) 256x256

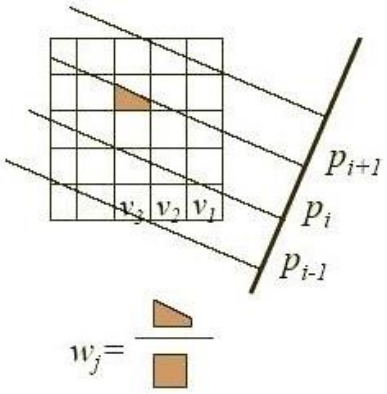


Figure. 4. Displays the projected data

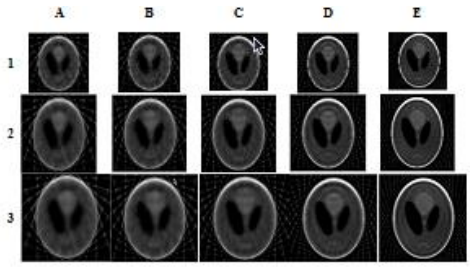


Figure. 5. The projection of Shepp Logan Phanton Image. Rows 1, 2 and 3 refer to the 64x64, 128x128, 256x256 data respectively. Columns A, B, C, D and E refer to the 10, 12, 15, 20 and 30 projection taken in 18, 15, 12, 9 and 6 angles respectively.

```
art()
{
   if ( not yet reached all the projections)
   {
        for (all elements in the projection)
        {
   calculate the value by multiplying
   the vector and the calculated data
   in the corresponding
   projection
        }
        calculate the error by subtracting
        the measured data from the calculated
        value
        for(all the rows)
        {
             correct the error by multiplying
     the difference with the
     calculated data.
             apply the corrected value to the
             vector.
        }
   }
   recursively call art function for
   remaining projections
}
```

Figure. 6. Art algorithm

are obtained using Radon function available in MATLAB. figure 4 shows the projection of the ray passed at a specific angle. The projection of a two dimensional function $f(x,y)$ is a set of line integrals Eqn. (1). The $f(x,y)$ is transferred to a row vector. The rays p_i passed at a specified angle

```
part()
{
   if ( not yet reached all the
                              projections)
   {

        omp_set_num_threads(number_of_threa
                                        ds);
        #pragma omp parallel for
           shared(elements) private(index)
            schedule(dynamic, num_elements)

        for (all elements in the
                              projection)
        {

            omp_set_num_threads(number_of_
                                threads);
            #pragma omp parallel for
              shared(elements) private(index)
              schedule(dynamic, num_elements)
              reduction(+:calculated value)

                 calculate the value by
                 multiplying the vector and
                 the calculated data in the
                 corresponding projection
        }

        calculate the error by subtracting
        the measured data from the calculated
        value

        omp_set_num_threads(number_of_threads
                                            );

        #pragma omp parallel for
          shared(elements) private(index)
          schedule(dynamic,num_elements)

        for(all the rows)
        {
            correct the error by multiplying
            the difference with the
            calculated data.

            apply the correction.
        }
   }

   recursively call part function for
                         remaining projections
}
```

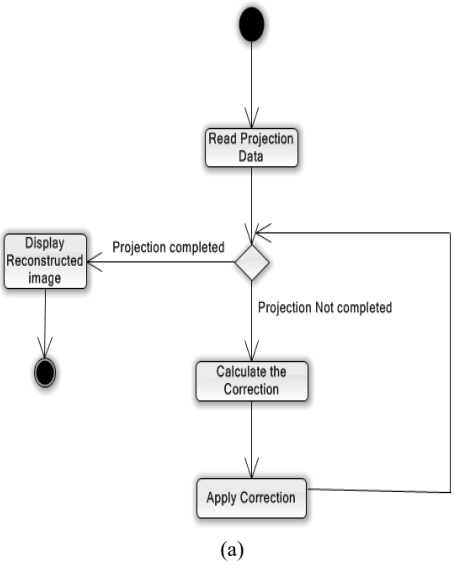
Figure. 7. pArt algorithm

collects data by calculating the weight matrix.

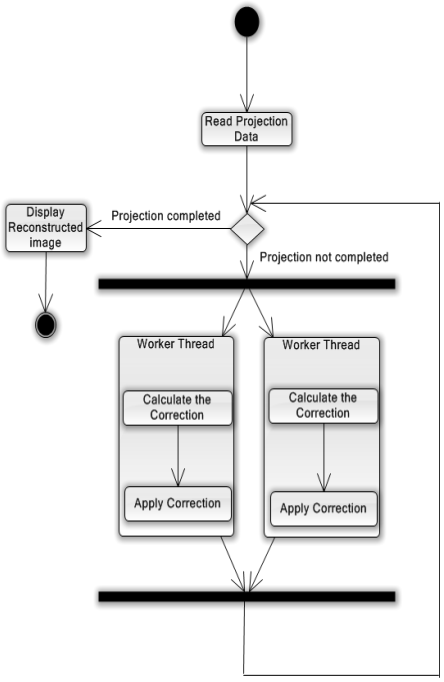
The projections of the Shepp Logan phantom in various angles are plotted in the figure 5. This is obtained by using *radon* function in matlab passing at which specific angles the object should be rotated. This system uses five different angles, such as 6^0, 9^0, 12^0, 15^0, 18^0 obtaining 30, 20, 15, 12, 10 numbers of projections respectively. Rows 1, 2 and 3 of figure. 5 refer to the projections taken from the images sizes 64x64, 128x128 and 256x256 respectively. Columns A, B, C, D, E refer to the 10, 12, 15, 20 and 30 projections taken in 18, 15, 12, 9 and 6 angles respectively.

Pseudo Code

The pseudo code for ART algorithm implemented



(a)



(b)

Figure. 8. UML Activity for reconstructing an image using ART. (a) Sequentially (b) Parallel

in MEX function executed sequentially is given in figure 6.

The pseudo code for pART algorithm implemented in MEX function executed parallel shows in figure 7.

UML Diagram

The operation of ART in sequential and parallel is symbolised in the figure 8(a) and figure 8(b) respectively. Parallel activity is pictured as Fork and Join.

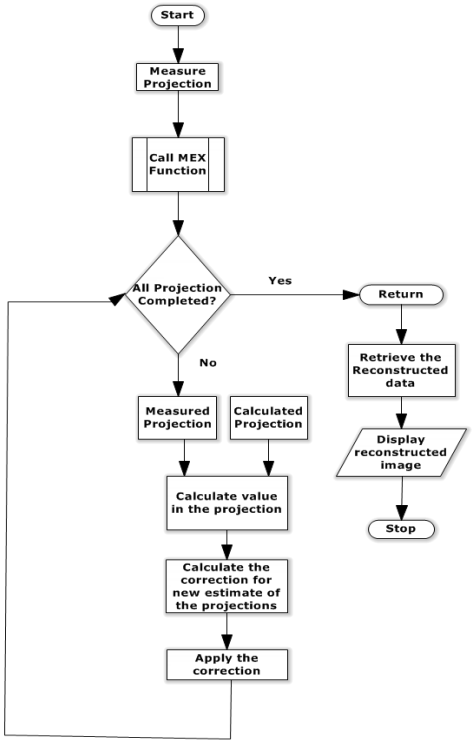


Figure 9: Flow chart for reconstructing an image using ART.

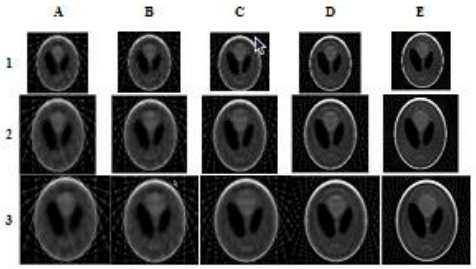


Figure. 10. The Reconstructed Shepp Logan Phantom. Rows 1, 2 and 3 refer to the 64x64, 128x128, 256x256 size of the Image respectively. Columns A, B, C, D and E refer to the reconstructed image from the 10, 12, 15, 20 and 30 projections of an image taken in 18, 15, 12, 9 and 6 angles in Sequential and parallel.

The data is read from the corresponding number of projections. This data is supplied into the MEX function to execute under single and multiple processors. For each projection the error value is calculated and the correction is applied

Initially, the program starts with initialization at each core. Then the calculated and measured projection data is co-distributed between the workers. After that, each worker calculates the correction and applies the correction as per the algorithm till all the projections are completed. Then the processed data is collected in a vector

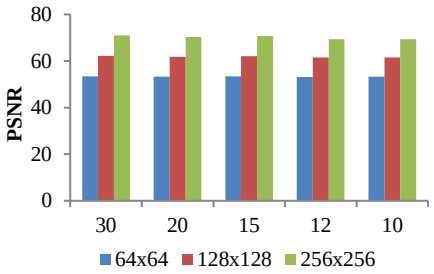


Figure. 11. Plotted the PSNR value obtained while reconstructing Shepp Logan Phantom Images on 64x64, 128x128 and 256x256 sizes using 30, 20, 15, 12 and 10 number of projections.

TABLE 1
PSNR VALUE IN DB FOR THE RECONSTRUCTED SHEPP LOGAN PHANTOM IN SINGLE CORE AND MULTI-CORE ENVIRONEMNT

Projections/ Sizes	30	20	15	12	10
64x64	53.3756	53.2965	53.4115	53.1949	53.2312
128x128	62.1416	61.7996	62.0032	61.4505	61.4595
256x256	71.0605	70.3131	70.6760	69.3787	69.3787

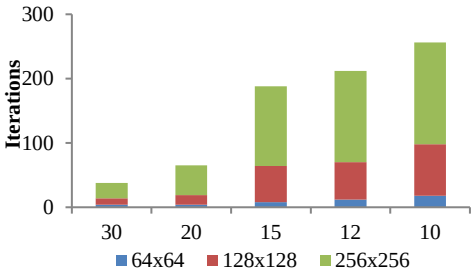


Figure. 12. Optimized number of iteration mandatory to reconstruct Shepp Logan Phantom Images on 64x64, 128x128 and 256x256 sizes using 30, 20, 15, 12 and 10 number of projections.

when the parallelism ends.

3. Results and Analysis

The results of constructing Shepp Logan Phantom image using ART in both sequential and parallel is given in figure 10. In this work, the time complexity of the phantom image of different size (64, 128 and 256) is compared in 2, 4 and 8 cores.

Peak-signal-to-noise ratio (PSNR) is used as a metric to check perceptual similarity between the original and reconstructed images. The PSNR value measured in db is tabulated in table 1. According to Chen et al (1998), PSNR above 40 db indicates a good perceptual fidelity. It can be observed that PSNR for the different size of images

TABLE 2
NUMBER OF ITERATIONS MADATORY TO RECONSTRUCT SHEPP LOGAN PHANTOM

Projections/ Sizes	30	20	15	12	10
64x64	4	4	8	12	18
128x128	10	15	56	58	80
256x256	24	46	124	142	158

TABLE 3
TIME COMPLEXITY OF RECONSTRUCTED PHANTOM IMAGE OF SIZE 64 X 64

Projections/ Cores	30	20	15	12	10
1 Core	1.6495	0.8426	1.3685	1.5327	3.017
2 Core	1.376	0.7431	1.1729	1.4561	2.5387
4 Core	0.9769	0.4959	0.9359	1.1558	1.8708
8 Core	0.7266	0.4828	0.6749	0.8876	1.4534

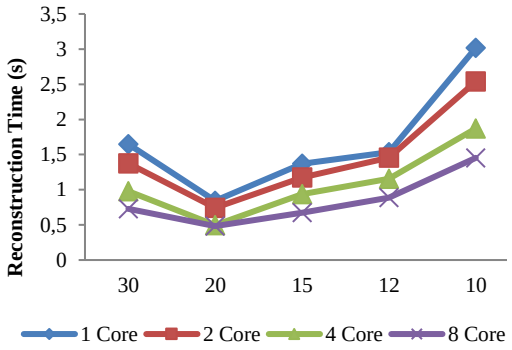


Figure. 13. A graph showing the Time Complexity of reconstructing Phantom image of size 64x 64 sequentially, parallel in 2, 4 and 8 core with respect to projections.

using various angles is above 60 db which indicates the excellent perceptual fidelity.

In figure 11 the PSNR value of the reconstructed image using ART for various sizes in different number of projections is graphed.

Reconstruction time taken by the Algebraic Reconstruction Technique for different size of phantom image in sequential and parallel using 2, 4 and 8 cores in an AMD Processor under LINUX platform. The time complexity of the reconstructed image of various sizes under 2, 4 and 8 cores is given with respect to the number of projections.

Reconstruction time taken by the Algebraic Reconstruction Technique for different size of phantom image in sequential and parallel using 2, 4 and 8 cores in an AMD Processor under LINUX platform. The time complexity of the reconstructed image of various sizes under 2, 4 and 8 cores is given with respect to the number of projections.

TABLE 4
TIME COMPLEXITY OF RECONSTRUCTED PHANTOM IMAGE OF SIZE 128 X 128

Projections / Cores	30	20	15	12
1 Core	28.8348	32.2032	80.0538	61.2674
2 Core	27.5855	29.6942	70.648	57.9245
4 Core	19.99	22.3794	51.1044	43.9571
8 Core	12.6774	14.533	31.1132	26.7882

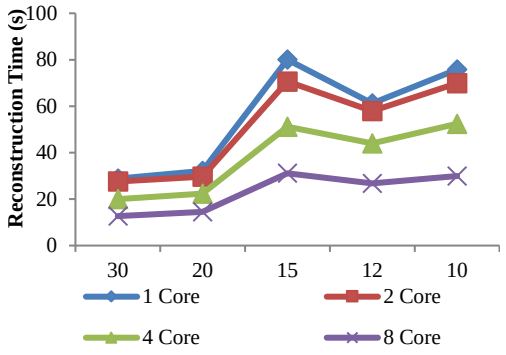


Figure. 14. A graph showing the Time Complexity of reconstructing Phantom image of size 128 x 128 sequentially, parallel in 2, 4 and 8 core with respect to projections

The optimized number of iteration to reconstruct an image in the three represented sizes at 30, 20, 15, 12 and 10 number of projections is tabulated in Table 2 and plotted in figure 12.

In figure 13, 14 and 15 the time complexity of phantom image of size 64, 128 and 256 reconstructed using 2, 4 and 8 cores with respect to 30, 20, 15, 12 and 10 is plotted respectively. Table 3, 4 and 5 tabulates the time complexity for 64, 128 and 256 size images respectively.

Table 3, 4 and 5 shows the reconstruction time taken by 1 Core (row1), 2 core (row2), 4 core (row 3), and 8 core (row 4) when using 30, 20, 15, 12 and 10 projections in the ART for image size 64, 128 and 256 respectively. It is observed that the time gradually reduces as the number of cores increases, for a given sets of projections. A graph is plotted to show the performance of the parallel system. It elucidates the time complexity of the system for a given number of projections using different cores of the parallel processor.

The time complexity of the system implementing parallel processor has got a considerable reduction of time consumptions which is certainly a high degree of utility to the user. A minor change in the time consumption will have a revolutionary impact while it is employed. The specific value of this finding is that the maximum number of core reconstruct the image is fast even for minimum number of projections.

TABLE 5
TIME COMPLEXITY OF RECONSTRUCTED PHANTOM IMAGE OF SIZE 256 X 256

Projections/ Cores	30	20	15	12	10
1 Core	715.2540	878.7210	1779.2300	1593.5500	1507.3800
2 Core	487.6430	673.5790	1330.1900	1107.0900	1048.2700
4 Core	380.0230	428.8580	947.2100	882.2380	785.8380
8 Core	218.3210	254.3550	503.6220	495.7050	414.3080

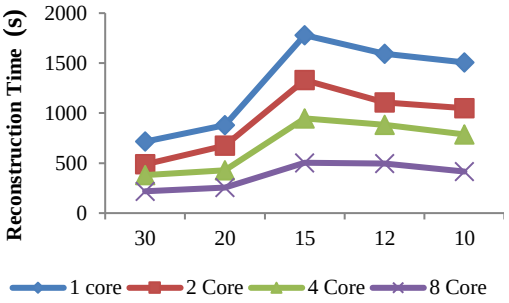


Figure. 15. A graph showing the Time Complexity of reconstructing Phantom image of size 256x256 parallel in 2, 4 and 8 core with respect to projections.

4. Conclusion

The number of iteration mandatory to reconstruct an image is optimized. The images are reconstructed sequentially as well as in parallel environment using different projection data sets. In this study, of Shepp Logan Phantom data is reconstructed using ART and pART. The results have shown encouraging indication of the efficiency of the parallelization of ART algorithm. In general, the pART algorithm gives a paramount computational efficiency better than ART. The computational efficiency of both ART and pART is reported in this article.

References

[1] R.Murugesan, M.Afeworki, J.A.Cook, N.Devasahayam, R.Tschudin, J.B.Mitchell, S.Subramanian, and M.C.Krishna. A broadband pulsed radio frequency electron paramagnetic resonance spectrometer for biological applications. Review of Scientific Instruments, 69(4), April 1998.

[2] K.Yamada, R.Murugesan, N.Devasahayam, J.A.Cook, J.B.Mitchell, S.Subramanian, and M.C.Krishna. Evaluation and comparision of

pulsed and continuous wave radiofrequency electron paramagnetic resonance techniques for in-vivo detection and imaging of free radicals. Journal of Magnetic Resonance, 154, 2002.

[3] R.A.Brooks and G.D.Chiro. *Theory of image reconstruction in computed tomography.* Radiology, 117, December 1975.

[4] G.L.Zeng. Image reconstruction a tutorial. *Computerized Medical Imaging and Graphics*, 25, 2001.

[5] P.F.C.Gilbert. Iterative Methods for the three dimensional reconstruction of an object from projections. *Journal of Theory, Biology*, 36, July 1972.

[6] P.M.V.Subbarao, P.Munshi, and K.Muralidhar. Performance of iterative tomographic algorithms applied to non destructive evaluation with limited data. *NDT and E International*, 30, 1997.

[7] C.N.Smith and A.D.Stevens. Reconstruction of images from radiofrequency electron paramagnetic resonance spectra. *The British Journal of Radiology*, 67, 1994.

[8] G.Placidi, M.Alecci, G.Gualtieri, and A.Sotgiu. Optimization of electron paramagnetic resonance image reconstructionusing filtered back projection followed by two dimensional deconvolution. *Journal of Magnetic Resonance*,A121, March 1996.

[9] G.Placidi, M.Alecci, and A.Sotgiu. Fourier reconstruction as a valid alternative to filtered back projection in iterative applications: Implementation of fourier spectral spatial epr imaging. *Journal of Magnetic Resonance*, 134, 1998.

[10] C.A.Johnson, J.A.Cook, D.McGarry, N.Devasahayam, J.B.Mitchell, S.Subramanian, and M.C.Krishna. Maximum entrophy reconstruction methods in electron paramagnetic resonance imaging. *Annals of Operations Research*, 119, January 2003.

[11] G. T. Herman and A. Lent. Iterative reconstruction algorithms. *Computers in Biology Medicine*, 6, January 1976.

[12] S.Sivakumar, Murali C.Krishna, R.Murugesan. *Evaluation of Algebraic Iterative Algoritms for Reconstruction of Electron Magnetic Resonance Images*, September 2010.

[13] Dan Gordon. Parallel ART for image reconstruction in CT using processor arrays. Department of Computer Science, University of Haifa, Haifa 31905, Israel. The International Journal of Parallel, Emergent and Distributed systems, Vol. 21, No. 5. October 2006, 365-380. January 2006.

[14] V. Kumar, A. Grama, A. Gupta, and G. Karaypis, "Introduction to parallel computing: Design and Analysis of Algorithms," *Redwood City, Calif.: Benjamin/Cummings*, 1994.

[15] Christopher D.Dharmaraj, Anthony R.Fletcher, Phuc N.Doan, Nallathamby Devasahayam, Shingo MAtsumato, Calvin A.Johnson, John A.Cook, James B.Mitchell, Sankaran Subramanian, Murali C.Krishna. Reconstruction for Time-Domain In-Vivo EPR 3-D Multi-Gradient Oximetric Imaging – A Parallel Processing Perspective. Radiation Biology Branch, Center for Cancer Research, National Cancer Institute, NIH, Bethesda, Marvland, USA, 1 June 2009.

[16] C.Terboven, T.Deselaers, C.Bischof and H.Ney, "Shared-memory parallelization for content-based image retrieval," in ECCV 2006 Workshop on Computation Intensive Methods for Computer Vision, 2006.

[17] Tiberius Duluman and Constantin Popa, Algebraic Reconstruction Technique versus Conjugate Gradient in Image Reconstruction from Projections, Proceedings of the Fifth Workshop on Mathematical Modelling of Environmental and Life Sciences Problems Constant, Romania, September, 2006, pp. 67–78

[18] Microslaw Miciak. Radon Transformation and Principle Component Analysis method applied in postal address recognition task. International Journal of Computer Science and Applications, Vol 7, No. 3, pp.33 – 44, 2010.

[19] S. Venturas, I. Flaounas, "Study of Radon Transformation and Application of its Inverse to NMR", Dept. of Informatics & Telecommunications, National and Kapodistrian University of Athens, Paper for "Algorithms in Molecular Biology" Course Assoc. Prof. I. Emiris, 4 July, 2005.

[20] S.Vandenberghe, Y.D.Asseler, R.Vandewalle. T.Kaupinen, M.Koole, L.Bouwens, K.Laere, I.Lemahieu, and R.A.Direckx. Iterative Reconstruction algorithms in nuclear medicine. *Computerized Medical Imaging and Graphics,* 25, 2001.

[21] Y. Censor and S. A. Zenios, Parallel Optimization: Theory, Algorithms, and Applications, Oxford University Press, New York, NY, USA, 1997.

[22] S. Kaczmarz, Angenaherte auflosung von systemen linearer gleichungen, Bulletin de

l'Academic Polonaise des Sciences et Lettres, A35 (1937) 355 -357.

[23] G. T. Herman, Image Reconstruction from Projections: The Fundamentals of Computerized Tomography, Academic Press, New York, NY, USA, 1980.

[24] P. P. B. Eggermont, G. T. Herman and A. Lent, Iterative algorithms for large partitioned linear systems, with applications to image reconstruction, Linear Algebra and Its Applications, 40 (1981), pp. 37–67.

[25] D.Raparia, J.Alessi and A.Kponou. *Algebraic Reconstruction Technique (ART)* AGS Department, Brookhaven National Lab, Upton, NY 11973, USA.Dan Gordon, 1998 IEEE.

[26] J. S. Kole and F. J. Beekman, "Parallel statistical image reconstruction for cone-beam X-ray CT on a shared memory computation platform," *Physics in Medicine & Biology*, vol. 50, pp. 1265–1272, 2005.

[27] C. Soviany, "Embedding data and task parallelism in image processing applications," Ph.D. dissertation, Technische Univ. Delft, 2003.

[28] Preeti kaur, Computer Science Department, Guru Nanak Dev University, Amritsar, Punjab, India, "Implementation of Image processing algorithms on the parallel platform using matlab", International Journal of Computer Science & Engineering Technology (IJCSET).7

Jurnal Ilmu Komputer dan Informasi (Journal of a Science and Information). 10/2 (2017), 96-101
DOI: http://dx.doi.org/10.21609/jiki.v10i2.459

A GOAL QUESTION METRIC (GQM) APPROACH FOR EVALUATING INTERACTION DESIGN PATTERNS IN DRAWING GAMES FOR PRESCHOOL CHILDREN

Dana Sulistiyo Kusumo, Mira Kania Sabariah, and Kemas Rahmat Saleh Wiharja

Software-Information and Data Engineering Research Group,School of Computing, Telkom University,
Jl. Telekomunikasi no. 1, Bandung, Indonesia

E-mail: danakusumo@telkomuniversity.ac.id, mirakania@telkomuniversity.ac.id,
bagindokemas@telkomuniversity.ac.id

Abstract

In recent years, there has been an increasing interest to use smart devices drawing games for educational benefit. In Indonesia, our government classifies children age four to six years old as preschool children. Not all preschool children can use drawing games easily. Further, drawing games may not fulfill all Indonesia's preschool children's drawing competencies. This research proposes to use Goal-Question Metric (GQM) to investigate and evaluate interaction design patterns of preschool children in order to achieve the drawing competencies for preschool children in two drawing Android-based games: Belajar Menggambar (in English: Learn to Draw) and Coret: Belajar Menggambar (in English: Scratch: Learn to Draw). We collected data from nine students of a preschool children education in a user research. The results show that GQM can assist to evaluate interaction design patterns in achieving the drawing competencies. Our approach can also yield interaction design patterns by comparing interaction design patterns in two drawing games used.

Keywords: *Interaction design pattern, goal-question metric, preschool children, drawing games*

Abstrak

Pada masa ini, terjadi peningkatan penggunaan aplikasi permainan menggambar pada perangkat pintar untuk tujuan pendidikan. Di Indonesia, pemerintah mengklasifikasikan anak-anak usia empat sampai dengan enam tahun sebagai anak usia dini. Namun tidak semua anak usia dini dapat menggunakan aplikasi permainan menggambar dengan mudah. Lebih lanjut lagi, aplikasi permainan menggambar tidak memenuhi semua kompetensi menggambar anak usai dini. Penelitian ini mengusulkan untuk menggunakan *Goal-Question Metric (GQM)* untuk menginvestigasi dan mengevaluasi pola desain interaksi anak usia dini dalam pencapaian kompetensi menggambar anak usia dini pada dua aplikasi menggambar berbasis Android: Belajar Menggambar dan Coret: Belajar Menggambar. Kami mengumpulkan data dari sembilan siswa Pendidikan Anak Usia Dini (PAUD) dalam sebuah *user research*. Hasil penelitian ini menunjukkan bahwa GQM dapat membantu untuk mengevaluasi pola desain interaksi dalam pencapaian kompetensi menggambar anak usia dini. Pendekatan yang kami usulkan dapat juga menghasilkan pola desain interaksi dengan membandingkan pola desain interaksi dalam dua aplikasi permainan menggambar yang digunakan.

Kata Kunci: *pola desain interaksi, goal-question metri, anak usia dini, permainan menggambar*

1. Introduction

Child computer interaction has been a fast-growing topic in the field of human computer interaction in recent years. Physically, four to six years old children could use wide range of gesture-based mobile application [1]. The government of Indonesia determines the education for four to six years old children as preschool education program. In recent years, the interest for creating drawing education application for children has increased. The drawing activity itself is full with stimulus for the growing process and development of the children. Beside that, the drawing activity gives visible and permanent traces that represents the characteristic of children. So it is very suitable to be done by the children [2]. By drawing, children could train the fine motor skills, creativity, imagination, concentration, memory, patience and the passion in learning. Drawing can also develop the children's capability in line with their talent and passion. And of course, by drawing, the children can know their surrounding world better [3].

Based on the previous empirical study, the children do not always understand about how to use tools and other movement in using a mobile application [4]. Therefore, the purpose of this research is to investigate the relationship between

children-game interaction and preschool children' drawing competencies. One of the main problems is the absence of the guidance in designing interaction patterns that is most suitable with the development stage of the children's fine motor skills. The interaction design pattern is a repeated solution for common usability problem in interface design [5]. The previous research [6] showed that interaction design pattern is one of the most important factor in designing a mobile application.

This research proposes a framework to analyze good interaction design pattern for drawing activities. We analyzed how children responded to several interaction design patterns and their interaction styles during drawing activities. After that, we defined the most proper interaction design pattern for children so that could meet the learning goal in preschool education program. Beside proposing the framework, we also assesed how the used interaction design patterns could meet the drawing competencies for children that is required by preschool education program. This assesment, in turn can provide information to the parents and the teachers of preschool children about how far an application can support children in mastering drawing competencies of preschool children. Interaction design patterns can be used by game developers as a guide to develop drawing game that will help the children in choosing the right features for drawing.

Related Work

Interaction design pattern is a design pattern for interaction between a user and a system [5]. Design pattern is defined as a repeatable solution for recurring software design problems [7]. Therefore, interaction design pattern is defined as a repeatable solution for recurring interaction problems. Interaction design patterns can be built from an existing application [8] following these processes: drafting existing interaction patterns, user research, testing interaction design pattern. Another approach using existing data to identify and analyze interaction design patterns use inductive and deductive approaches [9].

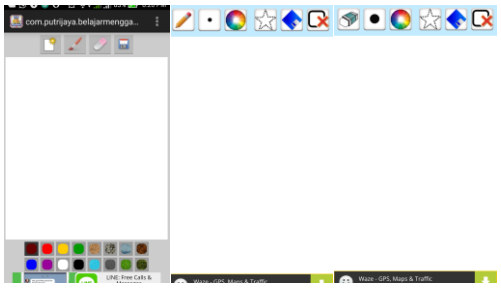
Goal Question Metric (GQM) is a software measurement framework for software products, processes and quality [10]. For measurement to be efective, we must have clear goals. Then, we detail goals into questions and develop metrics to answer the questions. So that, we can be confident that metrics used have strong relation with defined goals. GQM is used to detect design pattern using source code metrics combining with machine learning [11]. In recent literature, GQM has been used to guide a metric development for fulfilling security goal in the cloud [12], analysis of the use of the C preprocessor in community and industry [13] and designing a mobile app for behavior change [14]. GQM is also used to recommend design patterns for software designers and developers [15]. In the HCI field, GQM help to develope a usability metric for mobile application [16].

We propose to use an approach to analyze existing applications to identify and analyze interaction design patterns from [8]. Furthermore, we use GQM in order to ensure that interaction design patterns identified meeting drawing competencies for preschool children. In this research, we first use GQM to set goal of the competency achievement and then empirically identify and analyze interaction design patterns from existing applications.

2. Methods

This research proposes to use Goal-Question-Metric (GQM) [10] for measuring interaction design pattern and analysing interaction design pattern of preschool children in order to achieve drawing competencies for preschool children. In this research, we asked 9 preschool children to complete tasks in two drawing games and then we collected interaction data (detailed in Data collection sub-section). The detail of GQM for this research is described as follows: **Goals** (achieving drawing competencies for preschool children using interaction design patterns); **Question** (are drawing competencies included in drawing games that are observed to preschool children? and how long does it take for children to complete relevant tasks in drawing games?); **Metric** (to clasify and map drawing competencies into observed drawing games and to measure time of interaction design patterns used by preschool children for completing drawing tasks).

This research also used the two first steps of interaction design pattern construction framework proposed by Pauwels et al. [8] for evaluating interaction design pattern: 1) collect interaction design patterns form an existing application; 2) user research. The aim of the first step is to investigate existing interaction design patterns and to find their problems. The aim of the second steps is to analyze interaction problems and to propose solutions. In the user research, participants were asked to used two drawing game, and then we investigated their interaction design patterns. In Pauwels et al. [8], they only use one application, but in our research, we used two applications. The aim to use two drawing games is to compare and find similarities in and probelems of interaction design patterns between two draw-



Learn to draw | Scratch: Learn to draw (home) | Scratch: Learn to draw (eraser)

Figure 1. Two drawing games used in this research (Learn to Draw and Scratch: Learn to Draw)

TABLE 1
MAPING AMONG THE TASKS, COMPETENCIES AND FEATURES IN THE DRAWING GAMES

Task	Competencies	Features
Drawing	- Making vertical, horizontal, curved (left/right), tilt (left/right), and circle - Imitating shapes - Completing manipulative motion for drawing shapes - Classifying things according to shape, color and size	- Brush - Eraser
Color selection	- Completing manipulative motion for drawing shapes - Classifying things according to shape, color and size	- Brush - Eraser - Color selection
Eraser	The same as drawing task	Eraser

TABLE 2
PARTICIPANTS' USER PERSONAS

Factor	First persona	Second persona
Personality	A child frequently uses gadget, such as: smartphone.	A child infrequently uses gadget, such as: smartphone.
Goals	Fulfill drawing competencies using observed drawing games	
Behavior	Drawing using papers, pencils, erasers and manual drawing tools. A child usually draws on a paper sketching using a pencil then coloring using coloring tools.	
Knowledge	-Able to understand letters - Able to understand numbers - Able to understand simple shapes and figures - Able to understand colors	

TABLE 3
PRESCHOOL CHILDREN'S DRAWING COMPETENCIES AS GOALS MAPPED INTO DRAWING GAMES

Goals
Goals of common competencies for all drawing games
To make meaningful scratch To make vertical, horizontal, curved (left/right), tilt (left/right), and circle To draw human complete with limbs To draw and paint creatures lived and death
Goals of competencies specific to certain drawing games (not all games can provide these)
To thicken curved and straight lines To imitate the basic shapes (circle, triangle and rectangle) To imitate shapes To match, show and mention more than 11 colors
Goals of competencies that are not fully supported by any drawing games
To paint, draw, make patterns, sewing, weave, pierce, and formed with a variety of tools and materials

ing games used. In this research, we used two drawing games (Figure 1): Belajar Menggambar (in english: Learn to Draw) and Coret Belajar Menggambar (in english: Scratch: Learn to Draw) downloaded from Google PlayStore. These games were selected because their features are simple and easy to use. Therefore, we could concentrate on measuring interaction design patterns and duration of game played.

Data collection in this research used a user research aiming to identify context and problems of interaction design in the games. By completing the user research, we can expect frequently important problems found when a user play a drawing game. In this stage, research participants were asked to complete some task scenarios using two drawing games. These task scenario had been mapped to Indonesia's preschool children's drawing competencies (Table 1). The aims are to yield interaction design patterns supporting the preschool children's drawing competencies. For the purpose of this paper, we only limit to discuss interesting interaction design patterns resulted from our proposed approach.

Preschool children were asked to complete the same task scenarios in two drawing games, then we interviewed them to investigate processes and challenges of completion of the tasks. We measured time completion for each participant when completing the tasks. In the data analysis, we compared how the tasks' completion in two drawing games to indentify problems and proposed solutions so that we can propose interaction design patterns.

We conducted a user research in a Kindergaden of Raudlatul Athfal located in village of Sukapura, Dayeuhkolot, Bandung, Indonesia. The sample method was random. In the begining, we asked all students to participate in our user research, but only nine students agreed to join the user research. The children were asked to complete the task scenarios as aforementioned in the Research methodology section. We used a Samsung Galaxy Tab A. First, nine children were asked to complete all the tasks and instructions in the first drawing game, Learn to Draw. After finishing the first game, then the children were asked to complete the same activies in the second drawing game, Scratch: Learn to Draw. During completing the game, we observed and noted interaction between the participants and games. Finally, we in-

TABLE 4
TASKS' COMPLETION TIME FOR LEARN TO DRAW (IN SECOND)

	1	2	3	4	5	6	7	8	9
Drawing	<5	<5	<5	5-10	5-10	<5	<5	<5	<5
Eraser	<5	5-10	<5	5-10	5-10	5-10	<5	<5	<5
Color selection	<5	<5	<5	5-10	5-10	<5	<5	<5	<5

TABLE 5
TASKS' COMPLETION TIME FOR SCRATCH: LEARN TO DRAW (IN SECOND)

	1	2	3	4	5	6	7	8	9
Drawing	<5	<5	<5	<5	<5	<5	<5	<5	<5
Eraser	-	>10	-	-	-	-	-	-	-
Color Selection	<5	<5	<5	<5	<5	<5	<5	<5	<5

terviewed their experiences completing the tasks.

The participants in our research were children age four to six years old. Based on Sabariah et al. [4], we classified two user personas for preschool children (see Table 2). In our user research, three children were in the first persona and six children were in the second persona.

The goals of GQM in this research refered to the Indonesian curriculum of preschool children's drawing competencies. However, not all competencies can be mapped into the drawing games because there are activites in the competencies that can not be represented into digital applications. Goals that were expected to be achieved are listed in Table 3.

When comparing the same competencies for both drawing games, we only investigated the same tasks for both games as follows: Making vertical, horizontal, curved (left/right), tilt (left / right), and circle; Imitating shapes; Completing manipulative motion for drawing shapes; Classifying things according to shape, color and size.

Basically, these competencies represent common and specific competencies for drawing games (Table 3). Table 1 maps between the tasks and competencies, which are goals in GQM, for the drawing games.

Table 4 and Table 5 show completion time for each task in Table 1. Time was measured from time needed for each kid to complete the tasks after instructions given. For example: in Table 4, child 4 needed 5 to 10 seconds to complete drawing task after an instuction given.

3. Results and Analysis

We compared time measurement in Table 4 and Table 5. Then, we correlated and analyzed data qualitatively using contextual information gathered from the observation and interview in the user research. In this section, we also highlight identified interaction design patterns as results of our proposed method.

Goal-Question-Metric (GQM) was used to analyze interaction design patterns for supporting goals acheivement, which are preschool children's drawing competencies. Table 1 shows the category of preschool children's drawing competencies mapped into the tasks and supporting features in the drawing games as the result of breakdown goal into question and metric. Therefore, the drawing games have achieved goals of our proposed GQM which common preschool children's drawing competencies are available in the drawing games (Table 1). However, each of the games supports competencies specific to certain drawing games (Table 3). For example, the game of Scratch: Learn to Draw supports the competency of thickening straight and curved line, but this competency is not supported by the game of: Learn to Draw.

Based on, interaction design pattern measurement (during observation of the user research), the children did not find the games difficult to play. Average time to complete the tasks was between 5 and 10 seconds (Table 4 and 5). There are two explanation about this average time. The first is the role of first participant, named Jugjug, who has experience playing similar games. Then, the other participants followed Jugjug's interaction patterns using the games. The second is simple user interfaces so that the participants could easily learn to use and make adaption with provided menus and features. This was supported by facts that the participants were faster to complete drawing and color selection tasks in the second game compared to the first game because the participants had better understanding of mental model for the patterns of drawing games from the first game.

We find that a good interaction design pattern separates all buttons in the main interface (**the first interaction design pattern identified as the result of our proposed method**). This pattern is shown in the Draw to Learn game, which drawing, eraser and color selection are separated. Meanwhile, in Scratch: Draw to Learn game, eraser button is only active when drawing button is pressed (Figure 1). Not all participants could understand this feature in Scratch: Draw to Learn game, so that almost all participants chose to not finish the eraser task (Table 5) and prefered to choose to press new menu to erase existing picture. We suggest to separate all these buttons (**the second pattern**) in accordance with the mental model of a user reflecting a real world

condition which the functions of drawing tools and eraser are separated.

We propose to use two games in this research to identify and analyze interaction design patterns. Compared to [8], the use of two application has benefits of speeding up to study existing interaction design patterns, finding problems and proposing solutions. The use of two application is relevant to A/B testing [17] aiming to choose better system design. By comparing two applications, we can check whether interaction design patterns, including their problems and solutions, exist in both applications or not. If an interaction design pattern exist in both applications we can take the better interaction design pattern among the same interaction design pattern. This can strengthen an interaction design pattern resulted from the analysis because there are common problems and solutions related to interaction between an user and a system. But, if an interaction design pattern only exists in one of two applications, then we can choose this interaction design pattern as a proposed interaction design pattern. For example, the competency of thickening straight and curved line can only be found in the game of Scratch: Learn to Draw (**the third pattern**). Therefore, relevant interaction design pattern for this competency is taken from the game of Scratch: Learn to Draw. Based on our empirical results, the first and second identified interaction design patterns are suitable for preschool children.

4. Conclusion

The use of Goal-Question-Metric (GQM) can assist to identify, analyze and evaluate interaction design patterns in drawing games. The goal is to achieve preschool children's drawing competencies using interaction design patterns. By formulating questions and metrics derived from the goal, it can yield a mapping of drawing competencies into features, tasks and interaction design patterns of the observed drawing games. Based on this research's findings, using our approach can yield interacton patterns. By using two applications, we expects to get more inter-action design patterns compared to Pauwels et al. [8] approach, which uses an existing application to identify and analyze interaction design patterns. However, to get best interaction design pattern, we suggest to use all stages in Pauwels et al. [8] approach. Because we only used two applications, we only got limited interaction design patterns. Therefore, we suggest to compare more applications to yield more interaction design patterns. We suggest that parents and teachers of preschool students can use and combine more than one drawing game in order to achieve all preschool children's drawing competencies. In addition, this can also be an opportunity for Indonesian's game developers to produce drawing games fulfilling all preschool children's drawing competencies through the use of interaction design pattern.

Acknowledgment

This work is supported by Telkom University Internal Research Fund Scheme 2016. The authors also thank to Robby Daryodi, Dwitika Diah Pangestuti and Indah Mekar Sari for the discussion and providing the data.

References

[1] A. Hiniker *et al.*, "Touchscreen Prompts for Preschoolers: Designing Developmentally Appropriate Techniques for Teaching Young Children to Perform Gestures," in *Proceedings of the 14th International Conference on Interaction Design and Children*, New York, NY, USA, 2015, pp. 109–118.

[2] N. Scheuer, M. de la Cruz, and J. I. Pozo, "Children talk about learning to draw," *Eur. J. Psychol. Educ.*, vol. 17, no. 2, p. 101, Jun. 2002.

[3] M. Fadlillah;, "Desain Pembelajaran PAUD: Panduan untuk pendidik, mahasiswa & pengelola pendidikan anak usia dini," 2012. [Online]. Available: http://library.fip.uny.ac.id/opac/index.php?p=show_detail&id=6866. [Accessed: 08-Feb-2017].

[4] M. K. Sabariah, V. Effendy, and M. F. Ichsan, "Implementation of Hierarchical Task Analysis for User Interface Design in Drawing Application for Early Childhood Education," *J. Educ. Learn. EduLearn*, vol. 10, no. 2, pp. 159–166, May 2016.

[5] D. A. Norman and S. W. Draper, Eds., *User Centered System Design: New Perspectives on Human-computer Interaction*, 1 edition. Hillsdale, N.J.: CRC Press, 1986.

[6] F. B. Leandro, A. Solano, and C. A. Collazos, "Proposing Interaction Patterns for Designing Videogames Supported in Smartphones," in *Proceedings of the XV International Conference on Human Computer Interaction*, New York, NY, USA, 2014, p. 41:1–41:2.

[7] E. Gamma, R. Helm, R. Johnson, J. Vlissides, and G. Booch, *Design Patterns: Elements of Reusable Object-Oriented Software*, 1 edition. Reading, Mass.: Addison-Wesley Professional, 1994.

[8] S. L. Pauwels, C. Hübscher, J. A. Bargas-Avila, and K. Opwis, "Building an interaction design pattern language: A case study," *Comput. Hum. Behav.*, vol. 26, no. 3, pp. 452–463, May 2010.

[9] N. Schadewitz and T. Jachna, "Comparing inductive and deductive methodologies for design patterns identification and articulation," presented at the International Design Research Conference IADSR 2007 Emerging Trends in Design Research, Hong Kong, 2007.

[10] V. R. Basili, "Software Modeling and Measurement: The Goal/Question/Metric Paradigm," University of Maryland at College Park, College Park, MD, USA, 1992.

[11] S. Uchiyama, A. Kubo, H. Washizaki, and Y. Fukazawa, "Detecting Design Patterns in Object-Oriented Program Source Code by Using Metrics and Machine Learning," *J. Softw. Eng. Appl.*, vol. 07, no. 12, p. 983, Nov. 2014.

[12] F. Yahya, R. J. Walters, and G. B. Wills, "Using Goal-Question-Metric (GQM) Approach to Assess Security in Cloud Storage," in *Enterprise Security*, Springer, Cham, 2017, pp. 223–240.

[13] C. Hunsen *et al.*, "Preprocessor-based variability in open-source and industrial software systems: An empirical study," *Empir. Softw. Eng.*, vol. 21, no. 2, pp. 449–482, Apr. 2016.

[14] P. Tikka, B. Woldemicael, and H. Oinas-Kukkonen, "Building an App for Behavior Change: Case RightOnTime," in *Proceedings of the Fourth International Workshop on Behavior Change Support Systems (BCSS2016)*, Salzburg, Austria, 2016.

[15] F. Palma, H. Farzin, Y.-G. Guéhéneuc, and N. Moha, "Recommendation System for Design Patterns in Software Development: An DPR Overview," in *Proceedings of the Third International Workshop on Recommendation Systems for Software Engineering*, Piscataway, NJ, USA, 2012, pp. 1–5.

[16] A. Hussain and E. Ferneley, "Usability Metric for Mobile Application: A Goal Question Metric (GQM) Approach," in *Proceedings of the 10th International Conference on Information Integration and Web-based Applications & Services*, New York, NY, USA, 2008, pp. 567–570.

[17] B. Hanington and B. Martin, *Universal Methods of Design: 100 Ways to Research Complex Problems, Develop Innovative Ideas, and Design Effective Solutions*. Beverly, MA: Rockport Publishers, 2012.

Jurnal Ilmu Komputer dan Informasi (Journal of a Science and Information). 10/2 (2017), 102-107
DOI: http://dx.doi.org/10.21609/jiki.v10i2.480

SPARSE CODING-BASED METHOD COMPARISON FOR LAND-USE CLASSIFICATION

Dewa Made Sri Arsa[1], Grafika Jati[1], Yulistiyan Wardhana[1] and M. H. Hilman[2]

[1]Faculty of Computer Science, Universitas Indonesia, Kampus UI, Depok, 16424, Indonesia
[2]Melbourne School of Engineering, University of Melbourne, Building 173, Melbourne, VIC 3010, Australia

E-mail: dewa.made51@ui.ac.id, hilmanm@student.unimelb.edu.au

Abstract

Land-use classification utilize high-resolution remote sensing image. The image is utilized for improving the classification problem. Nonetheless, in other side, the problem becomes more challenging cause the image is too complex. We have to represent the image appropriately. On of the common method to deal with it is Bag of Visual Word (BOVW). The method needs a coding process to get the final data interpretation. There are many methods to do coding such as Hard Quantization Coding (HQ), Sparse Coding (SC), and Locality-constrained Linear Coding (LCC). However, that coding methods use a different assumption. Therefore, we have to compare the result of each coding method. The coding method affects classification accuracy. The best coding method will produce the better classification result. Dataset UC Merced consisted 21 classes is used in this research. The experiment result shows that LCC got better performance / accuracy than SC and HQ. LCC method got 86.48 % accuracy. Furthermore, LCC also got the best performance on various number of training data for each class.

Keywords: *Land-use classification, high-resolution remote sensing image, Bag of Visual Word (BOVW), Sparse Coding (SC), Hard Quantization Coding (HQ)*

Abstrak

Klasifikasi penggunaan lahan memanfaatkan gambar penginderaan jauh beresolusi tinggi. Citra digunakan untuk memperbaiki masalah klasifikasi. Meski begitu, di sisi lain, masalahnya menjadi lebih menantang karena gambarnya terlalu rumit. Kita harus mewakili gambar dengan tepat. Pada metode yang umum untuk mengatasinya adalah Bag of Visual Word (BOVW). Metode ini membutuhkan proses pengkodean untuk mendapatkan interpretasi data akhir. Ada banyak metode untuk melakukan pengkodean seperti Hard Quantization Coding (HQ), Sparse Coding (SC), dan Locality-constrained Linear Coding (LCC). Namun, metode pengkodean itu menggunakan asumsi yang berbeda. Oleh karena itu, kita harus membandingkan hasil setiap metode pengkodean. Metode pengkodean mempengaruhi akurasi klasifikasi. Metode pengkodean terbaik akan menghasilkan hasil klasifikasi yang lebih baik. Dataset UC Merced terdiri dari 21 kelas yang digunakan dalam penelitian ini. Hasil percobaan menunjukkan bahwa LCC memiliki kinerja / akurasi yang lebih baik daripada SC dan HQ. Metode LCC mendapat akurasi 86,48%. Selanjutnya, LCC juga mendapat performa terbaik pada berbagai jumlah data pelatihan untuk masing-masing kelas.

Kata Kunci: *Klasifikasi penggunaan lahan, citra penginderaan jauh beresolusi tinggi, Bag of Visual Word (BOVW), Sparse Coding (SC), Hard Quantization Coding (HQ)*

1. Introduction

Remote-sensing technique has been used as an effective tool to monitor Land-use and land-cover classification. Moreover, remote sensing technique is used to observe dynamic changing of a land [1-3]. Nowadays, single object classification and land classification research are progressive due to the better quality of remote sensing image [4-7].

Land-use-based classification uses image from remote sensing. The image is processed to extract information of land-use. On remote sensing, representation and efficient identification are still open problem and challenging. A lot of previous research used analytical approach, which focused on pixel- or object based classification. It extracted spectral, texture, and geometrical attributes [8-12]. Nevertheless, the attribute is only used in a certain environment so it just produce less data representation.

The recent years, Bag of Visual Words (BOVW) model is implemented to solve Land-use classification problems. It uses remote sensing image data [13]-[15]. Research [13] uses unsuper-vised-

feature-learning approach with Sparse Coding variant that is called Orthogonal Matching Pursuit (OMP-k). Research [14] uses combination of several features. The features was learned using clustering technique. The features are represented in histogram with linear weighting. Research [15] utilize derived method from Sparse Coding, Hard Assignment Vector Quantization. Moreover, research [16] employ Convolutional Neural Network (CNN)-based method named Gradient Boosting Random Convolutional Network (BGRCN). That method use Ensemble CNN which has high complexity of single CNN. Thus, the learning phase takes more time.

Coding as a learning feature and coding has many variations. The variation namely Hard Quantization (HQ), Soft Quantization (SQ), Sparse Coding (SC), Local Coordinate Coding (LCC) , Locality Constrained Linear Coding (LLC), Laplacian Sparse Coding (LSC), Over-complete Sparse Coding (OSC), Saliency Coding (SaC), Super-vector Coding (SV), and Improved Fisher Kernel (IFK) [17]. Each method has different complexity. Bag of Visual Words utilize the coding method to get the data representation.

Land-use classification research usually uses free dataset from UC Merced. The dataset has high degree of difficulty. The dataset has 21 classes of Land-use. This research will compare the performance of several coding methods especially SC, LCC, and HQ for Land-use classification.

The rest of the paper is organize as follows. In the section II, we present method. The section III, result, and analysis are presented. Moreover, we concluded this research in section IV. The last section is the references.

Literature Review

In this part will explain about SC method, HQ, and LCC.

Sparse Coding (SC)
SC method is a method develop from VQ method. SC is a L1-norm regularization for getting a small value that is not 0. Equation(1) shows the sparse coding method.

$$\min_{u,v} \sum_{m=1}^{M} \| x_m - u_m V \|^2 + \lambda | u_m | \qquad (1)$$

That depends on $\| v_k \| \leq 1, \ \forall k - 1, 2, ..., K$. X is a SIFT descriptor and V is a codebook from K clustering.

Locality-constrained Linear Coding (LLC)
The LLC method is initiated by fixing the LCC method, which has a weakness to high computational complexity. This method implements a locality. Therefore, it is important. As a result, The LLC's encoding formula becomes [20] showed in equation(2).

$$\min_{c} \| x_i - CR_i^T \|_2^2 + \lambda \| d_i \Theta R_i \|_2^2 \qquad (2)$$

With
$d_i = \exp\left[-\beta \left(\| x_i - c_1 \|_2^2 ... \| x_i - c_M \|_2^2 \right) \right]$
which is $|R_i||_1=1$.

Hard Quantization (HQ)
HQ method presents any local feature with a nearest visual word but only gives good performance when use many vocabularies[21].

$$R_i = \arg\min_{R_i} \| x_i - CR_i^T \|_2^2 \qquad (3)$$

With $Card(R_i) = 1$, $\| R_i \|_2 = 1$, and $R_i > 0$

Besides using different coding techniques, the complexity of each method is also different. The complexity is shown in Table 1. HQ has the highest complexity and SC has the lowest complexity.

TABLE 1
THE COMPLEXITY OF HQ, SC, AND LLC METHODS

Methods	Complexity
HQ	O(M)
SC	$O(M^2)$
LLC	$O(M+K^2)$

2. Methods

In this section, we will describe about the dataset and the method we used on the experiment. We are also present the experiment results and the analysis.

Dataset

To conduct the experiment, we chose to use UC Merced dataset. This dataset is a free data which can be downloaded in http://vision.ucmerced.edu/datasets/landuse.html. It needs to know that this dataset has 21 classes. Figure 2 shows the example of each classes in UC Merced dataset. Each

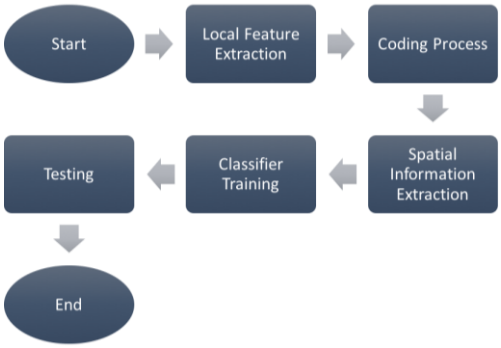


Figure 1. Research method

classes have 100 images, so there are 2100 images for the experiment.

Research Method

Our research method can be seen in Figure 1. There are 5 main process need to be done. The first process is local feature extraction. Then, the coding process is conducted to get the sparse representation of the local feature. After that, the spatial information is extracted from the data based on sparse features. Moreover, the process is continued to the fourth process, classifier training. The final process is testing the performance of our model. Each step will be described below.

Local Feature Extraction
Local feature, extracted from the raw data, is on the image patches form. To extract the local feature, we used Scaled Invariant Feature Transform (SIFT) method. We use dense SIFT to get all information from the data. We set some parameters to be fixed. They are patch size, descriptor degree, and grid spacing. The patch size is set into 16 x 16 px; descriptor has 8 degrees, and grid spacing is set by 8 px. The codebook is set by 1024. Output of this process is descriptor of each patches from the image. Base on this setting, first we extract all of the patches from image. Then, each patch is processed by compute the gradient magnitude.

Coding Process
In this research, we compare the performance of Sparse Coding, Locality constrained Linear Coding, and Hard Quantization method on the coding quality for feature representation in classification task, especially land use classification. The input is the descriptor result from the local feature extraction process. Each local feature will be mapped into sparse representation and locality. The sparse representation means approaching some values close to 0 so that only a few features are active, whereas locality will provide the feature representation in linear form. This locality makes the final features linearly separated.

Spatial Information Extraction
The result of coding process is a code of local feature for each patch. This result is lacked of spatial information. To address this problem, we used Spatial Pyramid Matching (SPM) method [18]. We divided the image into 3 types of region, 1x1, 2x2, and 4x4. In the 1x1 region, spatial information is extracted on hole image. In the 2x2 region, image will be divided into 4 regions, and 16 regions for 4x4 region type. The function of this division is to eliminate redundant coding features. The input to extract spatial information from the data is the result of the coding process. Then, the result of this partition will be made into one array 1 and the data is ready to be trained using a classifier.

Classifier Training
The classifier used to classify data is the Support Vector Machine (SVM) classifier. Research [13-15] also uses this method as a classification method. In addition, this method is chosen because it is able to maximize margin in the formation of decision boundary.

3. Results and Analysis

On the experiment, we measured the accuracy of the classifier. Then, we inspected the influence of data training number.

Classification Accuracy

In here, we divided the training and testing data with ratio 4:1 for each classes. The result can be seen in Figure 3. From this result, LLC performed better than SC and HQ. Because of we used linear classifier, this result proves that LLC has better performance to mapped the local feature into linear space.

The Effect of Amount of Train Data on Accuracy

To know the ability to represent the features in each coding method, the researcher conducted an experiment using different amounts of trainer data. The amount of training data used is 10, 20, 40, and 60. Figure 3 shows the graph of the resulting accuracy. HQ is not involved in comparisons due to HQ dependence on large vocabularies. From Figure 4 it can also be seen that using LLC as a feature encoding provides better accuracy than using the SC method.

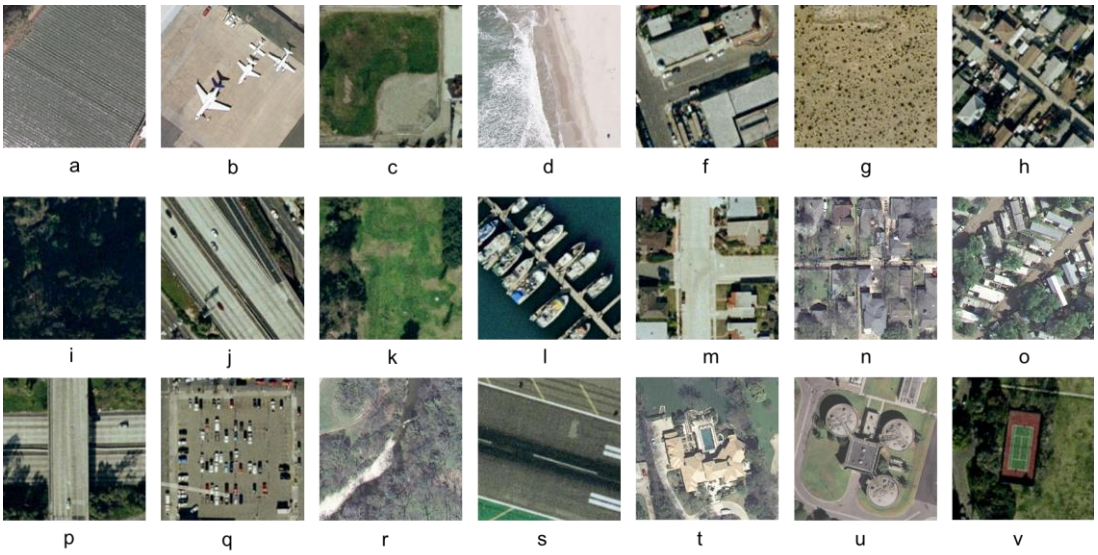


Figure 2. Example of each classes from UC Merced dataset. (a-v) agricultural, airplane, baseball diamond, beach, buildings, chaparral, dense residential, forest, freeway, golf course, harbor, intersection, medium residential, mobile homepark, overpass, parkinglot, river, runway, sparse residential, storage tanks, tennis court.

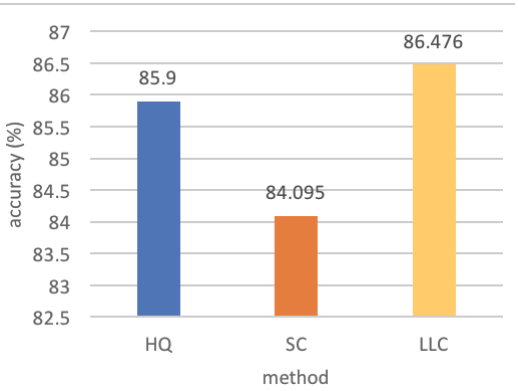


Figure 3. Accuracy of SC, LLC, and HQ

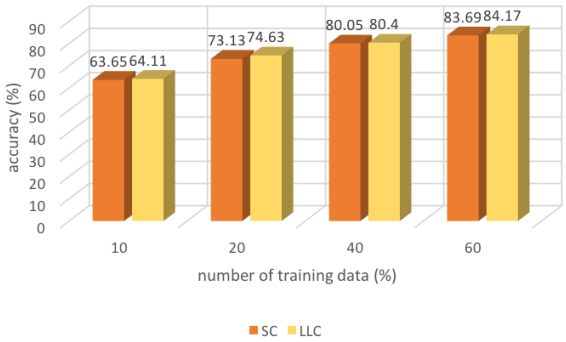


Figure 4. The accuracy of SC and LLC with variation of data training number for each class.

Analysis

From the measurement accuracy of the three methods, it can be seen that LLC has a better ability than SC or HQ. This proves that the locality that is carried by LLC is important so that it can represent better data. When it comes to land use classification, it relates to the amount of data that can be used as training data. LLC also shows its ability better than SC using little data. However when compared to [16], the accuracy of

LLC is lower. But high accuracy is followed by high complexity in model development.

4. Conclusion

This study has conducted a comparison between HQ, SC, and LLC methods. The measurement results show that LLC has better performance compared to HQ and SC. The number of training data used for the training also determines the accuracy. The more the number of train data, the more improved model recognition capabilities. The highest accuracy was obtained by LLC method of 86.476% for UC Merced dataset.

From the results of this study, it can be done further research which do boosting the method of coding to improve recognition performance. It can also inspect the possibility of other factors besides sparsity and locality that are important in the coding process.

References

[1] R. K. Jaiswal, R. Saxena, and S. Mukherjee, "Application of remote sensing technology for land use/land cover change analysis," *J. Indian Soc. Remote Sens.*, vol. 27, no. 2, pp. 123–128, Jun. 1999.

[2] J. Rogan and D. M. Chen, "Remote sensing technology for mapping and monitoring land-cover and land-use change," *Prog. Plann.*, vol. 61, no. 4, pp. 301–325, May 2004.

[3] Q. H. Weng, J. X. Zhang, P. Gamba, and G. Xian, "Foreword to the issue on remote sensing of regional land use and land cover," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 2, no. 2, pp. 50–53, Jun. 2009.

[4] F. Palsson, J. R. Sveinsson, J. A. Benediktsson, and H. Aanaes, "Classification of pansharpened urban satellite images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 1, pp. 281–297, Feb. 2012.

[5] A. A. Ursani, K. Kpalma, C. C. D. Lelong, and J. Ronsin, "Fusion of textural and spectral information for tree crop and other agricultural cover mapping with very-high resolution satellite images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 1, pp. 225–235, Feb. 2012.

[6] J. A. dos Santos, P. H. Gosselin, S. Philipp-Foliguet, R. D. Torres, and A. X. Falcao, "Interactive multiscale classification of high-resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 6, no. 4, pp. 2020–2034, Aug. 2013.

[7] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. 18th ACM SIGSPATIAL Int. Conf. Adv. Geogr. Inf. Syst.*, San Jose, CA, USA, 2010, pp. 270–279.

[8] M. Pesaresi and A. Gerhardinger, "Improved textural built-up presence index for automatic recognition of human settlements in arid regions with scattered vegetation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 4, no. 1, pp. 16–26, Mar. 2011.

[9] I. A. Rizvi and B. K. Mohan, "Object-based image analysis of highresolution satellite images using modified cloud basis function neural network and probabilistic relaxation labeling process," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 12, pp. 4815–4820, Dec. 2011.

[10] R. Bellens, S. Gautama, L. Martinez-Fonte, W. Philips, J. C.-W. Chan, and F. Canters, "Improved classification of VHR images of urban areas using directional morphological profiles," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 10, pp. 2803–2813, Oct. 2008.

[1]] A. K. Shackelford and C. H. Davis, "A combined fuzzy pixel-based and object-based approach for classification of high-resolution multispectral data over urban areas," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 10, pp. 2354–2363, Oct. 2003.

[12] P. Gamba, F. Dell'Acqua, G. Lisini, and G. Trianni, "Improved VHR urban area mapping exploiting object boundaries," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 8, pp. 2676–2682, Aug. 2007.

[13] A. M. Cheriyadat, "Unsupervised Feature Learning for Aerial Scene Classification," IEEE Trans. Geosci. Remote Sens., vol 52, no. 1, Jan. 2014.

[14] Li-Jun Zhao, P. Tang, and Lian-Zhi Huo, "Land-Use Scene Classification Using a Concentric Circle-Structured Multiscale Bag-of-Visual-Words Model," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 7, no. 12, Dec. 2014.

[15] S. Chen and YingLi Tian, "Pyramid of Spatial Relatons for Scene-Level Land use

Classification," IEEE Trans. Geosci. Remote Sens., vol 53, no.4, April 2015.

[16] F. Zhang, Bo Du, and L. Zhang,"Scene Classification via a Gradient Boosting Random Convolutional Network Framework," IEEE Trans. Geosci. Remote Sens., vol. 54, no. 3, March 2016.

[17] C. Wang and K. Huang, "How to use Bag of Words Model Better for Image Classification," Image and Vision COmputing, vol. 38, pp. 65-74, 2015

[18] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: spatial pyramid matching for recognizing natural scene categories, Computer Vision and, Pattern Recognition 2006, pp. 2169–2178.

[19] J. Yang, K. Yu, Y. Gong, T. Huang, Linear spatial pyramid matching using sparse coding for image classification, Computer Vision and Pattern Recognition 2009, pp. 1794–1801

[20] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, Y. Gong, Locality-constrained linear coding for image classification, Computer Vision and, Pattern Recognition 2010,pp. 3360–3367.

[21] G. Csurka, C.R. Dance, L. Fan, J. Willamowski, C. Bray, Visual categorization with bags of keypoints, European Conference on Computer Vision International Workshop on Statistical Learning in Computer Vision, 2004.

Jurnal Ilmu Komputer dan Informasi (Journal of a Science and Information). 10/2 (2017), 108-115
DOI: http://dx.doi.org/10.21609/jiki.v10i2.481

SUPERVISED MACHINE LEARNING MODEL FOR MICRORNA EXPRESSION DATA IN CANCER

Indra Waspada[1], Adi Wibowo[1], and Noel Segura Meraz[2]

[1]Department of Informatics, Faculty of Science and Mathematics, Diponegoro University, Tembalang, Semarang, 50275 Indonesia
[2]Department of Micro-Nano Mechanichal Science and Engineering, Nagoya University, Nagoya, 464048, Japan

E-mail: indrawaspada@undip.ac.id, bowo.adi@undip.ac.id , noel@robo.mein.nagoya-u.ac.jp

Abstract

The cancer cell gene expression data in general has a very large feature and requires analysis to find out which genes are strongly influencing the specific disease for diagnosis and drug discovery. In this paper several methods of supervised learning (decision tree, naïve bayes, neural network, and deep learning) are used to classify cancer cells based on the expression of the microRNA gene to obtain the best method that can be used for gene analysis. In this study there is no optimization and tuning of the algorithm to assess the fitness of algorithms. There are 1881 features of microRNA gene expresion, 22 cancer classes based on tissue location. A simple feature selection method is used to test the comparison of the algorithm. Expreriments were conducted with various scenarios to asses the accuracy of the classification.

Keywords: *Cancer, MicroRNA, classification, Decision Tree, Naïve Bayes, Neural Network, Deep Learning*

Abstrak

Data ekpresi gen sel kanker secara umum memiliki feature yang sangat banyak dan memerlukan analisa untuk mengetahui gen apa yang sangat berpengaruh terhadap spesifik penyakit untuk diagnosis dan juga penemuan obat. Pada tulisan ini beberapa metode supervised learning (decisien tree, naïve bayes, neural network, dan deep learning) digunakan untuk mengklasifikasi sel kanker berdasarkan ekpresi gen microRNA untuk mendapatkan metode terbaik yang dapat digunakan untuk analsisa gen. Dalam studi ini tidak ada optimasi dan tuning dari algoritma untuk menguji kemampuan algortima secara umum. Terdapat 1881 feature epresi gen microRNA pada 25 kelas kanker berdarkan lokasi tissue. Metode sederhana feature selection digunakan juga untuk menguji perbandingan algoritma tersebut. Exprerimen dilakukan dengan berbagai sekenario untuk menguji akurasi dari klasifikai.

Kata Kunci: *Kanker, MicroRNA, Klasifikasi, Decesion Tree, Naïve Bayes, Neural Network, Deep Learning*

1. Introduction

Cancer is the second deadliest disease after heart disease whith about 8.8 million cancer deaths by 2015. Moreover, one in six deaths is caused by cancer. The number of new cases are expected to increase by 70% over the next two decades [1]. It is generally recognized that cancer occurs due to gene abnormalities [2]. Gene's expression in the production rate of protein molecules are defined by genes [3]. Analyzing the gene expression profiles is the most fundamental approaches for understanding genetic abnormalities [4]. Micro Ribonucleic acid (microRNA) is known as one of the gene expressions that are very influential in cancer cells [5]. Gene's expression data, in general, has a very large number of features and requires analysis for diagnosis and disease analysis or to distinguish certain types of cancer and drug discovery [6].

Classification techniques of cancer cells based on gene expression data using machine learning methods have been developed rapidly in the analysis and diagnosis of cancer [7]. Classification techniques are definitely used to distinct the gene expression profiles for patients from cancer patients by type or even healthy patients [8]. One of the complicated problems in classification is to distinguish between different types of tumors (multiclass approach) which have a

very large quantities features of gene expression data [9]. For gene expression data, its high dimensionality and a relative fewer quantity numbers require much more consideration and specific preprocessing to deal with. In this case, it is important to aid users by suggesting which instances to inspect, especially for large datasets.

In constructing conventional machine learning systems require technical and domain skills to convert data into appropriate internal representations to detect patterns. Conventional techniques derive from single-spaced transformations that are often linear and limited in their ability to process natural data in their raw form [10]. Deep learning differs from traditional machines. In fact, in-depth learning allows a computational model consisting of several layers of processing based on neural networks to study data representation with varying levels of abstraction [10].

In this paper, the machine learning model has been implemented in studying features of genuine gene expression data and testing it in a classification model. We apply supervised learning in the form of a decision tree, naïve Bayes, and neural network compared with deep learning method in determining high-dimensional gene data pattern and achieving high accuracy. This comparison is intended to determine the reliability of the model tested in various cases, including feature selection.

The paper is structured as follows: Section 2 provides information on data and methods used for classification; Section 3 describes the results of a couple of methods from several scenarios of experiment and discussion. Finally section 4 the conclusion of paper and future works.

2. Method

Data sets

The datasets of MicroRNA expression in cancer and normal cell was occupied from National cancer institute GDC data Portal (**https://portal .gdc.cancer.gov/**). Table 1 shows the detail of datasets.

TABLE 1
SAMPLE NUMBER OF CANCER AND NORMAL CELL

Tissue	Cancer	Normal
Adrenal gland	259	3
Bile duct	36	9
Bladder	417	19
Brain	512	5
Breast	1096	104
Cervix	307	3
Colarectal	454	8
Esophagus	186	13
Head and neck	523	44
Kidney	544	71
Liver	372	50
Lung	519	46
Ovarium	489	0
Pancreas	178	4
Pleura	87	0
Prostate	497	52
Skin	97	2
Soft Tissue	259	0
Stomach	446	45
Thymus	124	2
Thyroid	506	59
Uterus	545	33

Decision Tree

Basically, the Decision Tree algorithm aims at obtaining a homogeneous subgroup of predefined class attributes by repeatedly repartitioning a heterogeneous sample group based on the value of the feature attribute [11], [12].

Next, divide the group into smaller and more homogeneous subgroups. Referring to the class attribute, the sample group partition is selected based on the feature attribute with the highest Information Gain value

The formula for calculating the information gain is derived from the following derivation [13]:

- Information expected to classify a tuple in D is expressed as:

$$Info(D) = -\sum_{i=1}^{m} pi \log_2(pi) \qquad (1)$$

with pi being the non zero probability that any tuple in D is part of class C_i and is estimated with $|C_{i,D}|/|D|$. The base log 2 function is used because the information is encoded in bits. Info (D) is the average amount of information needed to identify the Duplication class label D. Info (D) is also known as the entropy of D.

- The amount of information required on the classification is measured using the following formula:

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(Dj) \qquad (2)$$

The $\frac{|D_j|}{|D|}$ role as partition weight to j. $Info_A(D)$ is the information needed to classify the tuples of D based on A. The

smaller the information, the greater the purity of the partition.

- Information Gain is defined as the difference between the original information and the new information (obtained from the partition on A), so it can be formulated as follows:

$$Gain(A) = Info(D) - Info_A(D) \quad (3)$$

The iteration of the decision tree algorithm begins by partitioning the example using feature attributes with the largest Information Gain until it stops when the remaining value of the Information Gain attribute is below a certain threshold or the subgroup is homogeneous [11], [12]. In the end, it will produce a tree-like structure, with its branches being feature attributes and its leaves being subgroups. If there is an example as an input, then using the decision tree model that has been compiled it can be traced through the attribute of the input instance feature to predict the desired target attribute.

Naïve Bayes

A Naïve Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem (from Bayesian statistics) with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be independent feature model. In simple terms, a Naïve Bayes classifier assumes that the presence (or absence) of a particular feature of a class (i.e. attribute) is unrelated to the presence (or absence) of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 4 inches in diameter. Even if these features depend on each other or upon the existence of the other features, a Naïve Bayes classifier considers all of these properties to independently contribute to the probability that this fruit is an apple.

The advantage of the Naive Bayes classifier is that it only requires a small amount of training data to estimate the means and variances of the variables necessary for classification. Because independent variables are assumed, only the variances of the variables for each label need to be determined and not the entire covariance matrix.

Bayes is a conditional probability model for an example problem to be classified by the vector X = (x_1 ... ..x_n) with n example.

$$P(C_k|x_1 \dots x_n) \quad (5)$$

The problem with the above formula is that if the number of n is very large, it will need a very large range of values, so the probability becomes impossible. We have a tendency to do formulations on the model to provide additional use of Bayes theorem, its conditional probability is calculated as:

$$p(C_k|X) = p(C_k)p(C_k|X)/p(X) \quad (6)$$

The Bayesian probability terminology in the equation(6) can be written as Posterior = Likelihood / Evidence.

In practice, interest only exists in the numerator of the fraction, since the denominator is independent of C and the value of the given feature Fj, so the numerator is effectively constant. The numerator is equivalent to a joint probability model

$$p(C_k, x_1, \dots, x_n) \quad (7)$$

It can be rewritten as follows, by using chain rules for repeated applications on the definition of conditional probabilities as:

$$p(C_k, x_1, \dots, x_n) = p(C_k)p(x_1, \dots, x_n|C_k) \quad (8)$$

Recently the independent conditional Naive came into play: the assumption that each feature Fj is conditionally independent for every other Fi feature for j is not equal to I, given category C, this means that:

$$\begin{gathered} p(x_i|C_k, x_j) = p(x_i|C_k) \\ p(x_i|C_k, x_j, x_k) = p(x_i|C_k) = p(x_i|C_k) \\ p(x_i|C_k, x_j, x_k, x_i) = p(x_i|C_k) \end{gathered} \quad (9)$$

For i ≠ j, k, l then the combined model can be expressed as

$$\begin{gathered} p(C_{k,i}|x_1, \dots, x_j) \propto p(C_k, x_1, \dots, x_n) \propto \\ p(C_k)p(x_i|C_k)p(x_2|C_k) \\ p(x_3|C_k) \propto p(C_k) \prod \prod_{i=1}^{n} p(x_i|C_k) \end{gathered} \quad (10)$$

This means that based on the above independent assumption, the conditional distribution in the class C variable is:

$$p(C_k|x_i, \dots, x_j) = {}^1/_Z \, p(C_k) \prod \prod_{i=1}^{n} p(x_i|C_k) \quad (11)$$

where the evidence Z = p (x) is a scaling factor that depends on x1, ..., xn. That is constant if the value of feature variable is known.

Neural Network

Rapidminer provides neural network operator. The operator uses feedforward neural network algorithm with backpropagation algorithm for the training. Neural networks are inspired by biological neural networks, which are then developed as mathematical models. The structure of artificial neural networks consists of connected neurons that can process and transmit information.

One of the advantages of neural network is its adaptability that can change the structure of external and internal information obtained during the learning phase. The current use of neural networks is to find patterns from a set of data or to find complex models of relationships between inputs and outputs.

In the feedforward neural network, the information moves forward, one direction from the input to the output (via a hidden node) without the loop.

While backpropagation neural network (BPNN) algorithm uses to do looping at two stages of propagation and repeated, until achieved acceptable results (good). In this algorithm the error function (obtained from the output value compared to the correct answer) is fed back to the network as a reference to reduce the previous error value. Because the process of reduction is small for each stage it is necessary to do many training cycles until it reaches a small error value until it can be declared that it has reached the target.

Initially BPNN will look for an error between the original output and the desired output.

$$E_p = \sum_{i=1}^{j} (e_i) \qquad (12)$$

Where e is a nonlinear error signal. P shows pole to P; J is the number of units of output. The gradient descent method is shown in equation(13),

$$w_{k,i} = \mu \frac{\partial E_p}{\partial w_{k,i}} \qquad (13)$$

Back Propagation counts errors in the output layer σj, and hidden layer. Σj using equation(14) and equation(15):

$$\partial_l = \mu(d_i - y_i) f'(y_i) \qquad (13)$$

$$\partial_l = \mu \sum_i \partial_1 w_{l,} f'(y_i) \qquad (14)$$

Error in back propagation is used to update on weights and biases on output and hidden layers. Weight, Wij and bias, bj, then adjusted using the following equation:

$$w_{i,j}(k+1) = w_{i,j}(k) + \mu \partial_j y_i \qquad (15)$$

$$w_{l,j}(k+1) = w_{l,j}(k) + \mu \partial_j y_l \qquad (16)$$

$$b_j(k+1) = b_i(k) + \mu \partial_j \qquad (17)$$

Where, k is the epoch number and μ is the learning rate

Multi Layer Perceptron (MLP) was introduced to enhance the feed-forward with the mapping data set input to output. The structure of the MLP Algorithm consists of multiple node layers with a directional graph that each layer is fully connected to the next layer. Each node (other than the input node) is a neuron equipped with a nonlinear activation function. Multi Layer Perceptron utilizes back-propagation method in its training phase. The arrangement of MLP consists of several layers of computing units that implement sigmoid activation functions, and are linked to each other by feed-forward.

Deep Learning

Deep Learning is based on a multi-layer feed-forward artificial neural network that is trained with stochastic gradient descent using back-propagation. The network can contain a large number of hidden layers consisting of neurons with tanh, rectifier and maxout activation functions. Advanced features such as adaptive learning rate, rate annealing, momentum training, dropout and L1 or L2 regularization enable high predictive accuracy. Each compute node trains a copy of the global model parameters on its local data with multi-threading (asynchronously), and contributes periodically to the global model via model averaging across the network.

The operator starts a 1-node local H2O cluster and runs the algorithm on it. Although it uses one node, the execution is parallel. You can set the level of parallelism by changing the Settings/Preferences /General/Number of threads setting. By default, it uses the recommended number of threads for the system. Only one instance of the cluster is started and it remains running until you close RapidMiner Studio.

The Boltzmann engine is modeled with an input layer and a hidden layer that usually consists of binary units for each unit. The hidden layer is processed as stochastic (deterministic), recurrent (feed-forward). A generative model that can estimate distribution on observations for traditional

discriminative networks with labels. Energy on the network and Probability of a unit state (Scalar T expressed as temperature) is described as equation(18)

$$E(s) = -\sum_i a_i s_i - \sum_{i<j} s_j w_{i,j} s_i \quad (18)$$

A bipartite graph: No later-feed connection, feed-forward. Restricted Boltzmann Machine (RBM) has no T factor, the rest is similar to BM. An important feature of RBM is the visible unit and hidden unit are independent, which saves on good results later:

$$P(s_j = 1) = \frac{1}{1+e^{\left(-\frac{\Delta E}{T}\right)}} = \sigma\left(\left(s + \sum_{i=1}^{m} w_{i,j} s_i\right)/T\right) \quad (19)$$

$$P(v|h) = \prod_{i=1}^{m} p(v_i|h) \quad (20)$$

$$P(v|h) = \prod_{j=1}^{n} p(v_i|h) \quad (21)$$

Two characters used to define a Restricted Boltzmann Machine: The state of all units: obtained through the distribution of possibilities; Network weights: gained through training

As previously noted, RBM aims to estimate the distribution of input data. This goal is fully determined by weight and input. Energy defined for RBM is shown in equation(22):

$$E(v,h) = -\sum_i a_i v_i \sum_j b_j h_j - \sum_j \sum_i h_j w_{i,j} v_i \quad (22)$$

Distribution on the visible layer on RBM:

$$P(v) = \frac{1}{z}\sum_h e^{-E(v,h)} \quad (23)$$

Where, Z is a partition function defined as the sum of all possible configurations (v, h)

Training for RBM: Maximum Likelihood learns probability against vector x with parameter W (weight) is:

$$P(v) = \frac{1}{z}\sum_h e^{-E(v,h)} \quad (24)$$

$$P(x;W) = 1/Z(W)\, e^{-E(x;W)}$$

$$Z(W) = \sum_x e^{-E(x;W)} \quad (25)$$

$$P(x;W) = 1/Z(W)\, e^{-E(x;W)} \quad (26)$$

$$Z(W) = \sum_x e^{-E(x;W)} \quad (27)$$

3. Results and Analysis

The experiment purpose is to compare the performance of several supervised machine learning methods. In determining which method is best, the performance of the method is checked by evaluating the accuracy of the results. Classification accuracy is calculated by determining the percentage of tuples placed in the correct class. We compute the class precision, class recall and accuracy of the method defined as

$$Precision = \frac{tp}{tp+fp} \quad (22)$$

$$Recall = \frac{tp}{tp+fn} \quad (23)$$

$$Accuracy = \frac{tp+tn}{tp+tn+fp+fn} \quad (24)$$

where *tp* (true positive) is a properly classified positive example, *tn* (true negative) is a correctly classified negative example, *fn* (false negative) is a incorrectly classified positive example and *fp* (false positive) is a incorrectly classified negative example

In the first scenario, all classes of cancer were tried to classify according to 1881 features of microRNA. The normal class is a combination of all normal cell samples from different types of tissue. Based on figure 1 shown that deep learning method is very stable to classify multiclass for the precision value due to the ability of deep multi layer on deep learning are able to give optimal weight of each feature for multiclass case. Similar result shown on the class recall results as can be seen in Figure 2. Moreoover, deep learning method is able to get the recall class value> 60%.

The accuracy result of each algorithm obtained for this first scenario are; Deep learning 91.49%; Naive bayes 61.54%; Decision tree 34.15%; Neural network 5.48%. Based on these results shows that deep learning has the highest accuracy, while the neural network is very small. Neural networks are implemented with a total of 50 iterations to reduce computational time as result the weighting of neurons is unoptimal.

In the second scenario, normal and cancer of breast cells were tested for classification with 1881 microRNA features. Based on figure 3 shows that class precision of deep learning has the highest True Positive value at 100%. Moreover, according to Figure 4, only deep learning method which has achievement balanced of recall class between cancer and normal. In addition, the accuracy value, deep learning is superior compare

to other methods with accuracy 99.12%; While other methods are as follows: naïve bayes 90.35%; Decision tree 96.49%; Neural network 91.23%.

In the third scenario, a simple feature selection (expression value> 10,000) is tested on normal and cancer breast cells classification. Feature selection reduce the microRNA feature number to 3 (has-mir-10b, 21, 22). Based on figure 5 shows deep learning and neural network have the similar performance in precision, moreover other methods correspondingly have high precision value. The similar result is also perceived in the recall value as shown in figure 6.

In the fourth scenario, normal and cancer of breast cells are tested for classification with selected microRNA features according to the diagnostic criteria (has-mir-10b, 125-b1,125b-2, 141, 145, 155, 191, 200a, 200b , 200c, 203a, 203b, 21,210, 30a, 92a-1, 92a-2). Bsaed on figure 7 shows that deep learning, decision tree, and neural network have a high precision results. As same as the recall according to figure 8, deep learning and neural network have high recall achievement with 100%. Moreover, the accuracy value of each method are; deep learning 100%; Naïve bayes 93.86%; Decision tree 99.12%; neural network 100%.

In the fifth scenario, normal and cancers of cervix cells are tested for classification with 1881 microRNA features. Based on figure 9 shows that nearly all methods can have high precision results, except True Negative on neural networks. The identical results shows for recall according to figure 10.

In the sixth scenario, normal and cancer of

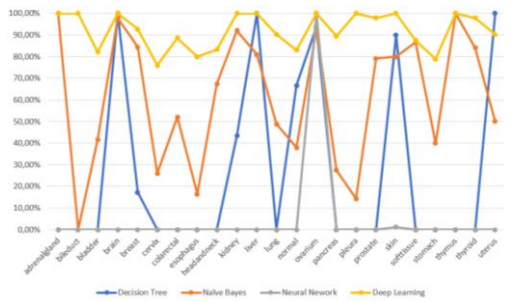


Figure. 1. Class Precision of multi classes cancer.

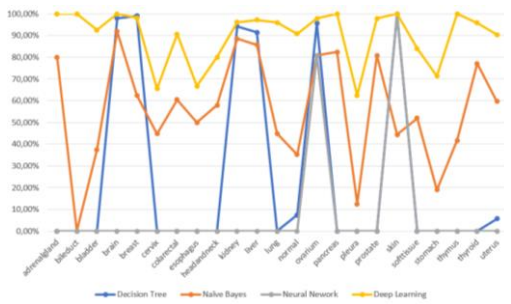


Figure. 2. Class Recall of multi classes cancer.

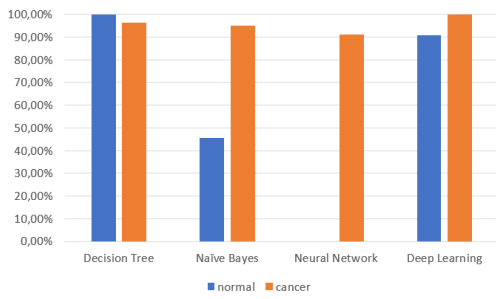


Figure. 3. Class Precision of breast tissue between normal and cancer cell all feature

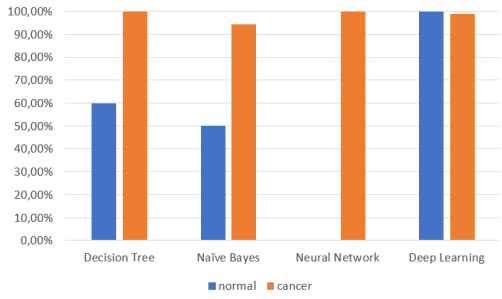


Figure. 4. Class Recall of breast tissue between normal and cancer cell all feature

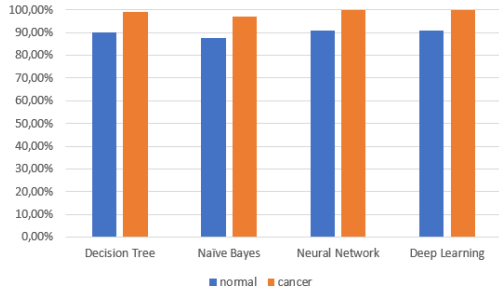


Figure. 5. Class Precision of breast tissue between normal and cancer cell with feature selection on criteria > 10.000

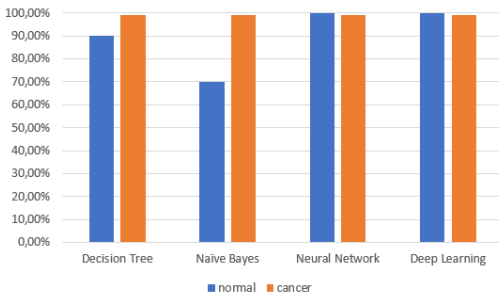


Figure. 6. Class Recall of breast tissue between normal and cancer cell with feature selection on criteria > 10.000

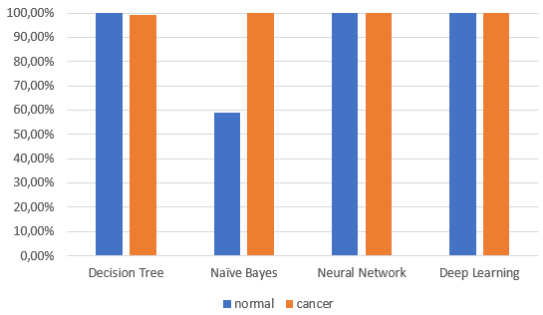


Figure. 7. Class Precision of breast tissue between normal and cancer cell with feature selection on diagnostic criteria (mir-21,)

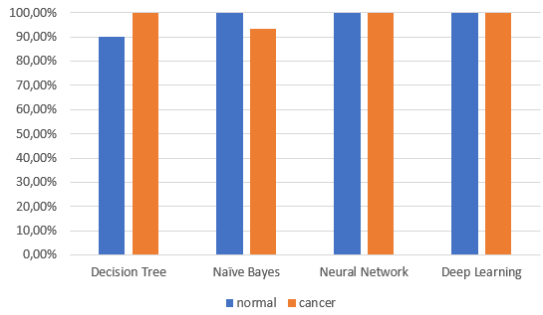


Figure. 8. Class Recall of breast tissue between normal and cancer cell with feature selection on diagnostic criteria (mir-21,)

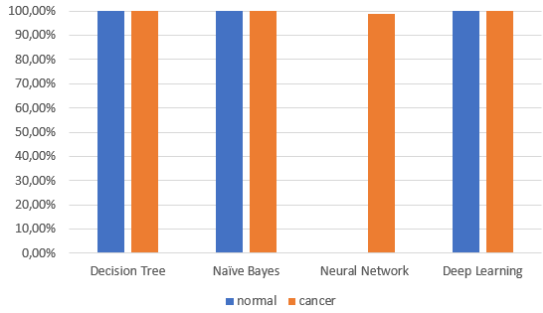


Figure. 9. Class Precision of cervix tissue between normal and cancer cell all feature

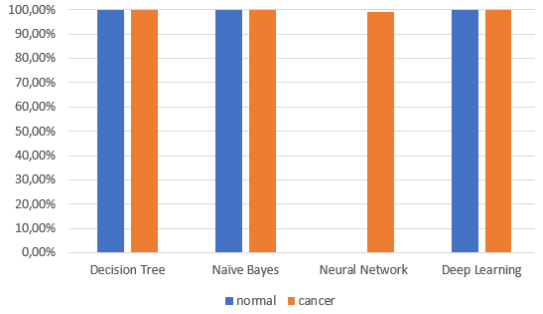


Figure. 10. Class Recall of cervix tissue between normal and cancer cell all feature

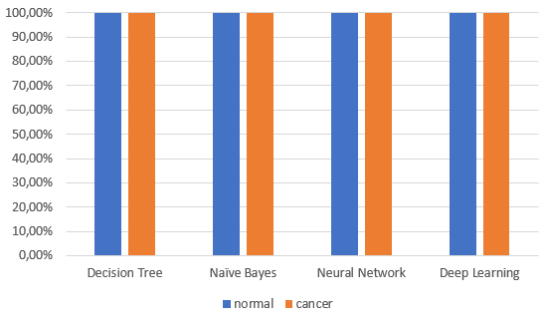


Figure. 11. Class Precision of cervix tissue between normal and cancer cell with feature criteria > 10.000

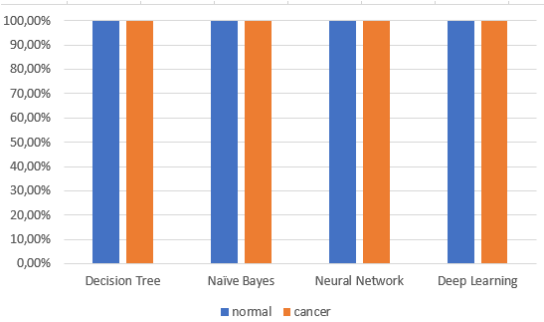


Figure. 12. Class Recall of cervix tissue between normal and cancer cell with feature criteria > 10.000

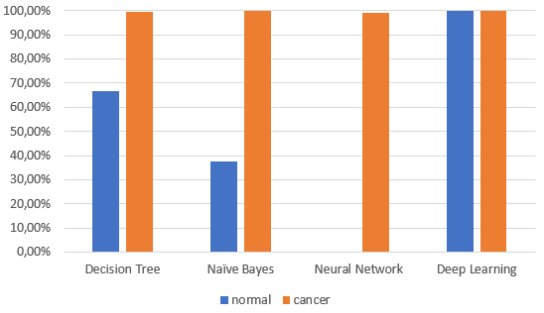


Figure. 13. Class Precision of cervix tissue between normal and cancer cell with feature criteria diagnostic

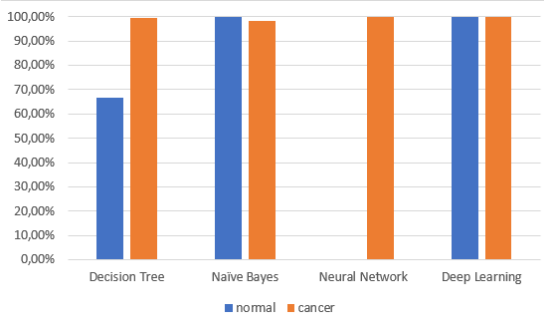


Figure. 14. Class recall of cervix tissue between normal and cancer cell with feature criteria diagnostic

cervical cells are tested for classification by simple feature selection (expression value > 10,000) and obtain the feature (has-mir-103a-1,103a-2,10b, 143,21,22). Based on figure 11 shows that all methods can have a perfect classification result. The equivalent results shown for recall according to figure 12.

In the last scenario, normal and cancer cervix cells are tested for classification by choosing diagnostic features with features (has-mir-146a, 155,196a-1,196a-2, 203a, 203b, 21, 221, 271, 27a, 34a). Based on figure 13 shows that only deep learning have a faultless classification result. The similar results shows in figure 14 for recall.

4. Conclusion

In this paper we have presented the performance of supervised machine learning method for classification of cancer cell expression gene data. Experimental results with various scenarios, all classes, breast classes, cervical classes, and some feature selection show that deep learning method is superior to decision tree, naïve bayes and neural network methods.

Acknowledgement

This work was supported by the RISTEKDIKTI, The Republic of Indonesia. Funding Source Number: 345-21/UN7.5.1/PP/2017.

References

[1] W. M. Centre, “Cancer,” 2017. .

[2] C. Lengauer, K. W. Kinzler, and B. Vogelstein, “Genetic instabilities in human cancers.,” *Nature*, vol. 396, no. 6712, pp. 643–649, 1998.

[3] D. J. Lockhart and E. A. Winzeler, “Genomics, gene expression and DNA arrays.,” *Nature*, vol. 405, no. 6788, pp. 827–36, 2000.

[4] S. R. Poort, F. R. Rosendaal, P. H. Reitsma, and R. M. BERTINA, “A common genetic variation in the 3'-untranslated region of the prothrombin gene is associated with elevated plasma prothrombin levels and an increase in venous thrombosis,” *Blood*, vol. 88, no. 10, pp. 3698–3703, 1996.

[5] H. Lan, H. Lu, X. Wang, and H. Jin, “MicroRNAs as potential biomarkers in cancer: Opportunities and challenges,” *Biomed Res. Int.*, vol. 2015, 2015.

[6] R. K. Singh and M. Sivabalakrishnan, “Feature selection of gene expression data for cancer classification: A review,” *Procedia Comput. Sci.*, vol. 50, pp. 52–57, 2015.

[7] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, “Machine learning applications in cancer prognosis and prediction,” *Comput. Struct. Biotechnol. J.*, vol. 13, pp. 8–17, 2015.

[8] H. Chen, H. Zhao, J. Shen, R. Zhou, and Q. Zhou, “Supervised Machine Learning Model for High Dimensional Gene Data in Colon Cancer Detection,” *2015 IEEE Int. Congr. Big Data*, pp. 134–141, 2015.

[9] F. Liao, H. Xu, N. Torrey, P. Road, and L. Jolla, “Multiclass cancer classification based on gene expression comparison,” vol. 2, no. 74, pp. 477–496, 2015.

[10] Y. LeCun, Y. Bengio, G. Hinton, L. Y., B. Y., and H. G., “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[11] H.-M. Lee and C.-C. Hsu, “A new model for concept classification based on linear threshold unit and decision tree,” in *Proceedings of the International Joint Conference on Neural Networks (IJCNN-90-Wash D.C. IEEE/INNS)*, 1990, pp. 631–634.

[12] Q. J. Ross, *C4.5 : programs for machine learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers, 1993.

[13] J. Han, M. Kamber, and J. Pei, *Data Mining Concepts and Techniques*, Third Edit. Morgan Kaufmann Publishers, 2012.