2020 The 12th International Conference on Future Computer and Communication (ICFCC 2020)

Selected, peer reviewed papers from the
2020 The 12th International Conference on Future Computer and Communication
(ICFCC 2020)
February 26-28, 2020, Yangon (Rangoon), Myanmar (Burma)

Edited by
Prof. Yutaka Ishibashi

The SCIence and Engineering Institute (SCIEI)
2448 Desire Avenue
Rowland Heights
LA
CA 91748
U.S.A
E-mail: info@sciei.org
Web: http://www.sciei.org/

ISBN 978-981-14-4787-7

Distributed *worldwide by*

The SCIence and Engineering Institute (SCIEI)
2448 Desire Avenue
Rowland Heights
LA
CA 91748
U.S.A
E-mail: info@sciei.org
Web: http://www.sciei.org/

Preface

Following the success of 2019, 2020 The 12th International Conference on Future Computer and Communication (ICFCC 2020) is to be held in conjunction with the 18th International Conference on Computer Applications (ICCA 2020) once again in Yangon (Rangoon), Myanmar (Burma) during February 26 through 28, 2020, which is organized by SCIence and Engineering Institute, USA, co-organized by University of Computer Studies, Yangon under Ministry of Education, Yangon (Rangoon), Myanmar (Burma), and technically supported by Nagoya Institute of Technology, Japan. The conference has arranged three days program: One day for registration, and Two days for Keynote and invited speeches, and oral presentations.

All accepted paper by 2020 The 12th International Conference on Future Computer and Communication (ICFCC 2020) conference proceedings would be published by WCSE on WCSE conference proceedings.

The scientific program consisted of 25 parallel oral and 2 poster sessions and included topics of Computer Engineering and Communication Engineering, The 3-day conference had about 150 participants. One day for registration, one day for keynote speech and invited talks, Prof. Nobuo Funabiki, Okayama University, Japan, Prof. Chih-Peng Fan, National Chung Hsing University, Taiwan, and Prof. Yu-Cheng Fan, National Taipei University of Technology, Taiwan, these three speakers from ICFCC 2020 have given their speeches. The contents of many of presenters can be found in the present contributions, arranged according to the above topics and the time sequence.

We gratefully acknowledge the help from University of Computer Studies, Yangon (Rangoon), Myanmar (Burma), excellent review works from technical committees, and hard onsite session chair works for 27 session chairs.

ICFCC 2020 Conference Chair

Prof. Yutaka Ishibashi, Nagoya Institute of Technology, Japan

Conference Committees

Conference Co-Chairs

Mie Mie Thet Thwin, University of Computer Studies, Yangon, Myanmar

Wen-Chung Kao, National Taiwan Normal University, Taiwan

Yutaka Ishibashi, Nagoya Institute of Technology, Japan

TPC Co-Chairs

Shinji Sugawara, Chiba Institute of Technology, Japan

Khin Mar Soe, University of Computer Studies, Yangon, Myanmar

Publicity Co-Chairs

Moe Pwint, University of Computer Studies, Mandalay, Myanmar

Saw Sanda Aye, University of Information Technology, Myanmar

Hiroaki Nishino, Oita University, Japan

Local Co-Chairs

Nang Saing Moon Kham, University of Computer Studies, Yangon, Myanmar

Khin Than Mya, University of Computer Studies, Yangon, Myanmar

TPC Members

Masaki Aida, Tokyo Metropolitan University, Japan

Chaodit Aswakul, Chulalongkorn University, Thailand

Chih-Peng Fan, National Chung Hsing University, Taiwan

Yu-Cheng Fan, National Taipei University of Technology, Taiwan

Toshiaki Fujii, Nagoya University, Japan

Akihiro Fujihara, Chiba Institute of Technology, Japan

Hiroshi Fujinoki, Southern Illinois University Edwardsville, USA

Masaru Fukushi, Yamaguchi University, Japan

Dai Hanawa, Nagoya City University, Japan

Takanori Hayashi, Hiroshima Institute of Technology,

Naohira Hayashibara, Kyoto Sangyo University, Japan

Takefumi Hiraguri, Nippon Institute of Technology, Japan

Pingguo Huang, Seijoh University, Japan

Takayuki Ito, Nagoya Institute of Technology, Japan

Yasunori Kawai, National Institute of Technology, Ishikawa College, Japan

Jong-Won Kim, Gwangju Institute of Science and Technology, Korea

George Kokkonis, University of Western Macedonia, Greece

Ryogo Kubo, Keio University, Japan

Teck Chaw Ling, University of Malaya, Malaysia

Takahiro Matsuda, Tokyo Metropolitan University, Japan
Aung Htein Maw, University of Information Technology, Myanmar
Takanori Miyoshi, Nagaoka University of Technology, Japan
Tom Murase, Nagoya University, Japan
Kentaro Nishimori, Niigata University, Japan
Hiroaki Nishino, Oita University, Japan
Thi Thi Soe Nyunt, University of Computer Studies, Yangon, Myanmar
Toshiro Nunome, Nagoya Institute of Technology, Japan
Hitoshi Ohnishi, The Open University of Japan, Japan
Takashi Okuda, Aichi Prefectural University, Japan
Kenko Ota, Nippon Institute of Technology, Japan
Takanobu Otsuka, Nagoya Institute of Technology, Japan
Kostas E. Psannis, University of Macedonia, Greece
Myint Myint Sein, University of Computer Studies, Yangon, Myanmar
Shun Shiramatsu, Nagoya Institute of Technology, Japan
Yuichiro Tateiwa, Nagoya Institute of Technology, Japan
Aye Thida, University of Computer Studies, Mandalay, Myanmar
Mie Mie Su Thwin, University of Computer Studies, Yangon, Myanmar
Masato Tsuru, Kyushu Institute of Technology, Japan
Hitoshi Watanabe, Tokyo University of Science, Japan
Shingo Yamaguchi, Yamaguchi University, Japan
Tatsuya Yamazaki, Niigata University, Japan
Kyoko Yamori, Asahi University, Japan
Hideaki Yoshino, Nippon Institute of Technology, Japan
Cheng Zhang, Waseda University, Japan
Thi Thi Zin, University of Miyazaki, Japan
Sutep Tongngam, National Institute of Development Administration (NIDA), Thailand
Janusz R. Getta, University of Wollongong, Australia
Bambang Leo Handoko, Bina Nusantara University, Indonesia
Remedios de Dios Bulos, De La Salle University, Philippines
May Aye Khine, University of Computer Studies, Yangon, Myanmar
Khin Mo Mo Tun, University of Computer Studies, Yangon, Myanmar
Tammam Tillo, Libera Universita di Bolzano-Bozen, Italy
Nwe Nwe Myint Thein, University of Information Technology, Myanmar
Kiattisak Maichalernnukul, Rangsit University, Thailand

Table of Contents

Chapter 1- Data Analysis and Intelligent Computing

Sentiment Analysis System for Myanmar News using K Nearest Neighbor and Naïve Bayes…….....1
Thein Yu, Khin Thandar Nwet

A Novel Clustering-based Class-asociation Rule Mining Method for Handling Class-Imbalanced Datasets...6
Tien-Dung Phan, Thanh-Tho Quan, Thi-Kim-Anh Vo

Neighbor Search with Hash Map Indexing Technique for Complex Networks……..……...……...11
Wai Mar Hlaing, Myint Myint Sein

Analysis of Outlier Detection on Structured Data……...….16
Khin Myo Myat, Si Si Mar Win

Real-time Big Data Analytics for Feature Selection on Apache Spark……...….22
Lwin May Thant, Sabai Phyu

Comparative Results of Dependent and Independent Variables Focused on Regression Analysis Using Test-Driven Development…….27
Myint Myint Moe, Khine Khine Oo

Chapter 2- Artificial Intelligence and Machine Learning

Statistical Machine Translation between Myanmar and Myeik……36
Thazin Myint Oo, Ye Kyaw Thu, Khin Mar Soe, Thepchai Supnithi

ICD-10 Auto-coding System Using Deep Learning…….46
Ssu-Ming Wang, Feipei Lai, Chang-Sung Sung, Yang Chen

Innovation Security of Beaufort Cipher by Stream Cipher Using Myanmar-Vigenere Table and Unicode Table……..52
Htet Htet Naing, Zin May Aye

Development of Spoken Language Recognition System for Humanoid Robot……...……57
Khaing Yee Mone, Yoshio Yamamoto

Building Large Scale Text Corpus for Joint Word Segmentation and Part-of-Speech Tagging of Myanmar Language…….……..63
Dim Lam Cing, Khin Mar Soe

A Fear State Judgement System for Alleviating Fear of Heights Gradually in VR…….68
Iku Kitanosono, Toshiyuki Haramaki, Hiroaki Nishino

Chapter 3- Image Processing

Persistent Operation of OF@TEIN+ Playground Verified by SmartX Multi-View Leveraged Visualization73
Muhammad Ahmad Rathore, SeungHyung Lee, JongWon Kim

Effective Multi-View for Human Activity Recognition on Skeletal Model…….79
Sandar Win, Thin Lai Lai Thein

Compressed Sensing Image De-noising Algorithm Based on L1-L2 Norm Regularization……84
Liu Ziming, Fang Changjie

QoE Comparison of AL-FEC Algorithms on H.265/HEVC Video and Audio Transmission with MMT…...90
Toshiro Nunome, Koki Makino

Chapter 4- Communication and Information System

Projected Iterative MVDR Beamforming for Null Broadening and First Sidelobe Suppression in the Presence of DOA Mismatch…….95
Raungrong Suleesathira

Bilateral Tele-Rehabilitation System with Electrical Stimulation by Using Cloud Service……...103
Yasunori Kawai, Koudai Houga, Hiroyuki Kawai, Takanori Miyoshi

Influences of Network Delay and Moving Velocity on Virtual Cooperative Work with Haptic Sense…….108
May Zin Oo, Yutaka Ishibashi, Khin Than Mya

An Investigation on Stability and Operability in Haptic Communication Systems……114
Hitoshi Watanabe, Pingguo Huang, Yutaka Ishibashi

Artificial Intelligence based 6G Intelligent IOT: Unfolding an Analytical Concept for Future Hybrid Communication Systems……..…122
Muhammad Haroon Siddiqui, Kiran Khurshid, Imran Rashid, Adnan Ahmed Khan

A Study On Deep Learning Based Real Time Road information Monitoring Device for Emergency Vehicle Guidance……130
Seona Park, Junghan Ha, Muwook Pyeon, Wonwoo Jung

Chapter 5- Advanced Information Technology and Management

A Study on Community Overlapping Detection Algorithms in Social Networks……..136
Eaint Mon Win, May Aye Khine

Ensemble Learning Method for Enhancing Healthcare Classification……...141
Pau Suan Mung, Sabai Phyu

A CSP-based Approach to Design a Subnet Solving a Network Construction Exercise for Beginners…….146
Yuichiro Tateiwa, Yoshifumi Hisanaga

GPS Trajectory Cleaning For Driving Behaviour Detection System…….....151
Tin Lai Lai Mon

Secure Healthcare System using Blockchain Technology……..156
HtweHtwePyone, KhinThan Mya

The Analysis of Landslide Based on Geographic Information System in Mon State, Myanmar……...162
Chaw Chaw Khaing, Thin Lai Lai Thein

Service Management Strategy for CDN……168
Xinhua E, Binjie Zhu, Hui Zhang, Yanjun Shi

A Content Sharing System Using Dynamic Fog Consisting of Peer-to-Peer Terminals and Its Simple Evaluation……...173
Takuya Itokazu, Shinji Sugawara

Chapter 1
Data Analysis and Intelligent Computing

ISBN 978-981-14-4787-7
Proceedings of 2020 the 10th International Workshop on Computer Science and Engineering (WCSE 2020)
Yangon (Rangoon), Myanmar (Burma), February 26- February 28, 2020, pp. 1-5

Sentiment Analysis System for Myanmar News using K Nearest Neighbor and Naïve Bayes

Thein Yu [1+] and Khin Thandar Nwet[2]

[1] University of Computer Studies, Yangon, Myanmar
[2] University of Information Technology, Myanmar

Abstract. With the explosive growth of internet technology, there are very large amount of information on the web for the internet users. Users not only use that information but also provide opinions for decision making process. Sentiment analysis or opinion mining is one of text classification techniques that identify and extract opinion described in a piece of text. Our aims in this paper are to develop automatic sentiment analysis system for Myanmar news and to annotate sentiment news. Therefore, this system creates sentiment annotated corpus for Myanmar news. Feature extraction and selection are very important for sentiment analysis to get higher performance. N-grams, Countvectorizer, and TF-IDF are used for feature selection and feature extraction. In this system, Myanmar news sentiment analysis system is implemented by using K Nearest Neighbor (KNN) and Naïve Bayes machine learning algorithms.

Keywords: Sentiment analysis, Naïve Bayes, K Nearest Neighbor, N-gram, TF-IDF.

1. Introduction

Opinion mining or sentiment analysis is a popular research field in the combination of information retrieval (IR) and natural language processing (NLP) and have common characteristics with other disciplines such as text mining and information extraction. Sentiment analysis or opinion mining is also the task that intends to refer the sentiment orientation in a document. Opinion mining is a technique of text mining that provides a way for individual and corporation to exploit the large amount of information available on the internet and to detect and extract polarity orientation of subjective information in text documents. There are three levels of sentiment: (i) document-based level; (ii) sentence-based level; and (iii) aspect-based level. In document-based and sentence-based sentiment analysis, it is implicitly assumed that the analysed document or sentence only discusses a single object. In general, sentiment analysis determines the sentiment orientation of a writer about some aspect or the overall contextual polarity of a document. Sentiment classification is a recent sub division of text classification which is concerned not only with the topic of document, but also with the expressed opinion. Sentiment classification also has different names, among which opinion mining, sentiment analysis, sentiment extraction, or affective rating. News can be good or bad, or seldom neutral. The statistical analysis of sentiment cues can give a powerful sense of how the latest news impacts important entities [1].

Sentiment analysis for Myanmar language has many challenges due to scarcity of resources such as automatic feature extraction tools, stemming, anaphora resolution, and name entity recognition etc. In this paper, sentiment analysis system of Myanmar news is proposed. Feature extraction and transformation are used in sentiment analysis to get better performance. N-Gram and TF-IDF are used in this system for feature extraction and transformation. K nearest neighbour and naïve bayes algorithms is applied in implementing sentiment analysis system.

[+] Corresponding author. Tel.: +959898681179
E-mail address: theinyu@ucsy.edu.mm.

The remaining parts of this paper are organized as follows; the related works are explained in section 2. The methodology is described in section 3. In section 4, experiment showed. Conclusion and future work is presented in section 5.

2. Related Work

Many sentiment analysis systems are existed for English language and other languages. Different methods are used with different resources at different levels for different systems.

Joseph Lilleberg, Yun Zxu, and Yanqing Zhang proposed a text classification system with semantics features. They used TF-IDF model and combined with word2vec model. They show performance result of TF-IDF, word2vec combined with wighted TF-IDF with stop words and without stop words[2]. In paper [3], authors implemented Chinese text classification system using N-gram based feature selection and text representation methods. They used four feature selection methods such as absolute text frequency, relative text frequency, absolute n-gram frequency and relative n-gram frequency and three text representation methods such as 0/1 logical value, n-gram frequency numeric value (TF) and TF-IDF value with support vector machine and naive bayes machine learning algorithm.

In paper [4], authors developed twitter sentiment system that use N-gram feature selection and combination. They use Sanders product review dataset. They model system using kern lab classifier, decision tree classifier, and naive bayes classifier. In paper [5], authors implemented improved text sentiment classification model that use TF-IDF and next word negation for feature. They used movie review dataset, product review dataset, and SMS spam dataset that are trained with linear support vector machine, maximum entropy, random forest, and multinomial naive bayes.

3. Proposed System

3.1. Preprocessing

Preprocessing is the important step in natural language processing. There are three preprocessing steps in the proposed system.

- Word Segmentation is the fundamental task in natural language processing that identify boundaries of word. Myanmar word segmentation is the process of placing spaces into textual data without other replacing or rewriting operations. This system used word segmentation tool from NLP Lab, University of Computer Studies,Yangon, Myanmar examples of segmented result are as follow:

  Sentence 1: သမ္မတ သည် သံဃာ့ မဟာ နာယက ဆရာတော်ကြီး များ ကို ဂါရဝပြု ပြီး ဩဝါဒ ခံယူသည်

- Tokenization is the process of separating up a sequence of strings into words, phrases, keywords and other elements. Tokens or words are separated and identified by white space, punctuation marks or line breaks.

- Stop words are commonly used words that are arranged to ignore for searching, retrieving, and other natural language processing tasks. Stop word removal is important in preprocessing step to get better performance result. Examples of Myanmar stop words are တွင်,နှင် ,များ,မှ,မှာ,က,ကာ,သော,၏,ပြီ,ပြီး,သည်,လျှင်,၌,၍,၏, and etc.

After removing stop words, sentence 1 may be as follow:

Sentence 1: သမ္မတ သံဃာ့ မဟာ နာယက ဆရာတော်ကြီး ဂါရဝပြု ဩဝါဒ ခံယူသည်

3.2. Feature Extraction and Transformation

- **N-gram :** N-gram is a language models that assign probabilities to the sequences of words. N-gram is based on bag of word model and has a sequence of word with n length. N-gram with length (n=1) is called unigram and length (n = 2) is also called bigram and then length (n = 3) is also called trigram. Text classification also depends on text representation to get higher accuracy [6].

Examples of n-gram words for sentence 1 after removing stop words is shown in table 1.

Sentence 1: သမ္မတ သံဃာ့ မဟာ နာယက ဆရာတော်ကြီး ဂါရဝပြု ဩဝါဒ ခံယူသည်.

Table 1: Examples of N-gram

N-gram	*Feature*
Unigram	'သမ္မတ', 'သံဃာ.', 'မဟာ','နာယက', 'ဆရာတော်ကြီး' ' ဂါရဝပြု' ' ကြဝါဒ' ' ခံယူသည်'
Bigram	'သမ္မတ သံဃာ', 'သံဃာ မဟာ့ 'မဟာ နာယက' ,'နာယက ဆရာတော်ကြီး' , ' ဆရာတော်ကြီး ဂါရဝပြု', ' ဂါရဝပြု ကြဝါဒ', ' ကြဝါဒ ခံယူသည်', ' ခံယူသည်'
Combination of Unigram and Bigram (Unigram + Bigram)	'သမ္မတ' , 'သမ္မတ သံဃာ.' 'သံဃာ့', သံဃာ. မဟာ' 'မဟာ' ,' မဟာ နာယက','နာယက' ,'နာယက ဆရာတော်ကြီး' ,' ဆရာတော်ကြီး' , 'ဆရာတော်ကြီး ဂါရဝပြု','ဂါရဝပြု', ' ဂါရဝပြု ကြဝါဒ', ' ကြဝါဒ, ' ကြဝါဒ ခံယူသည်' ,' ခံယူသည်'

- **CountVectorizer :** The Countvectorizer is one of the bag of word model to tokenize text document and build a vocabulary of known words. And then, it also encodes document into document term matrix vector with that vocabulary. The encoded term matrix vector contains length of entire vocabulary and integer count number of each word presented in the document. The encoded vector is transformed to array version of vector that can be directly used by machine learning algorithm.
- **TF-IDF Vectorizer:** TF-IDF is a term weighting method which shows the importance of a term in a document to present textual data. It compares the frequency of word described in an individual document as opposed to the entire document. TF-IDF is based on the bag-of-words (BoW) model and does not necessary to have position in text, semantics, co-occurrences in different documents, etc[7].

3.3. Machine Learning Algorithm.

- **Na ïve Bayes:** The bayesian classification is one of probabilistic learning method for text classification. Naive bayes classifier is an independent features model that the inclusion (or exclusion) of a particular feature of a class is unrelated to the inclusion (or exclusion) of any other feature[8]. The probability function of sentiment class given document is calculated using equation 1and 2.

$$P(c/d) = P(c)P(d/c)/P(d) \quad (1)$$

Where, c=sentiment class, p(c/d) = probability of sentiment class given document, p(d/c)= likelihood of document, p(d)= probability of document, p(c) =prior probability d=document
P(d) has the same value, so p(d)can be drop. Document can be presented as many features such as f_1, f_2, f_3,…f_n, function can be as follows:

$$P(c/d) = argmaxP(c) \prod_{f \epsilon F} P(f/c) \quad (2)$$

Where, f= features vector, c= sentiment class, d=document, p(f/c)= likelihood of features given to sentiment class

- **K Nearest Neighbor (KNN) Algorithm :**Nearest neighbor algorithm is one of the simplest machine learning algorithms. The KNN algorithm is also a lazy learning because the computation for the generation of the predictions is postponed until classification. The idea is to acquire the training set and then to predict the label of any new instance on the closest labeled neighbors in the training set. The algorithm works based on the rule that chooses minimum distance from the test data to the training samples to determine the K nearest neighbor. After defining K nearest neighbor, a simple majority of them is used to predict test data. The KNN works as follows: The distance between the test data and all the training samples are calculated. The distance may be calculated by any standard means. Euclidean distance is usually used. The K nearest neighbor may be considered if the distance of the training samples to the test samples is less than or equal to Kth smallest distance. The quality of the predictions relies on the distance measure. The KNN algorithm is suitable for applications for sufficient domain knowledge [9, 10].

4. Experiment

4.1. Experimental Apparatus

Naïve bayes and K nearest neighbor algorithms were experiment on the Myanmar News dataset. The dataset was split in the ratio of 80 % for training and 20 % for testing purposes. These experiments were carried out using an open source 'scikit-learn' Python library and NLTK Library on Jupyter Notebook.

4.2. Training Dataset

Myanmar news data are collected from web sites and ALT Tree bank for training and testing data. Sentiment corpus contains 2000 news for positive and 1000 news for negative. The task is to identify if positive or negative sentiment is expressed at document and sentence level. There has been found that many researchers used unbalanced dataset for sentiment analysis. Therefore, recently dataset are used in the proposed system.

4.3. Experimental Result

The hold out evaluation method is used in the experiment. Naïve bayes with countvectorizer (unigram) features gets greatest accuracy score values. The evaluation metrics such as accuracy, precision, recall and F1 measure were calculated using equation 3-6. Table 2 and 3 show performance results.

$$\text{Accuracy}=\frac{TP+TN}{TP+FP+TN+FN} \quad (3)$$

$$\text{Precision}=\frac{TP}{TP+FP} \quad (4)$$

$$\text{Recall}=\frac{TP}{TP+FN} \quad (5)$$

$$\text{F-Measure}=\frac{2*Precision*Recall}{Precision+Recall} \quad (6)$$

TP defines the number of positive news that are correctly classified, as positive,
FP is the number of negative news that are incorrectly classified as positive.
TN is the number of negative instances that are correctly classified as negative.
FN is the number of positive tuples that are incorrectly classified as negative news.

Table 2: Performance Results 1

Feature with TFIDF	**Naïve Bayes**	**KNN**
	Accuracy %	**Accuracy %**
Unigram	75.67	80.83
Bigram	69.33	74
Trigram	67.83	70
Unigram + Bigram	71.83	79.67
Bigram +Trigram	68.50	73.50
Unigram + Bigram+ Trigram	69.50	79.83

Table 3: Performance Results 2

Features with CountVectorizer	**Naïve Bayes**	**KNN**
	Accuracy %	**Accuracy %**
Unigram	82.33	68.67
Bigram	77.00	68.33
Trigram	69.50	68.33
Unigram + Bigram	81.17	68.17
Bigram+Trigram	74.17	68.33
Unigram + Bigram+ Trigram	79.83	68.00

5. Conclusion and Future Work

The proposed system is implemented to classify sentiment of news in Myanmar language. By using system, user can easily feel emotion of news. K nearest neighbor and naïve bayes machine learning algorithms are used in this system. This system compares accuracies of those algorithms with unigram, bigram, trigram, combination of unigram and bigram, combination of bigram and trigram (bigram+trigram), and combination of unigram, bigram, and trigram(unigram+ bigram + trigram) features. Naïve bayes with Countvectorizer(unigram) feature has highest accuracy value in the proposed system. In future, we intend to classify with many algorithms such as CNN, ANN, and RNN deep learning algorithms.

6. Acknowledgements

I specially thank to my supervisor, Dr Khin Thandar Nwet, Lecturer, University of Information Technology, Yangon, Myanmar. I would like to thank Dr Win Pa Pa, Professor, Natural Language Processing Lab, University of Computer Studies, Yangon, for word segmentation tool. I deeply thanks to anonymous reviewers for my paper.

7. References

[1] B. Liu. Sentiment Analysis and Opinion Mining", Synthesis Lecturer on Human Language Technologies, Morgan & Claypool Press, 2012

[2] J. Lilleberg. Y. Zhu, and Y. Zxang. Support Vector Machines and Word2vec for Text Classification with Semantic Features, Proc. 20151IEE 14th Inl'l Coni. on Cognitive Inlormatics & Cognitive Computing IIccrcn51, 2015

[3] Z. Wei, D. Maio, J. H. Chauchai, and R. Z. Wenli. N-gram based feature selection and text representation for Chinese text classification, International Journal of computational Intelligence Systems, Vol.2, No.4, pp- 365-374December, 2009

[4] P.B. Awachate and V.P. Jsgursagar. Improved Twitter Sentiment Analysis using N Gram feature selection and Combination, International Journal of Adavanced Research in computer and communication engineering, Vol.5Issue 9, September 2016.

[5] B. Das and S. Chakraborty. An Improved Text Sentiment Classification Model using TF-IDF and Next word Negation,

[6] Jurafsky ,D., & Martin, J.,H.(2018) .N-gram Language Models. Speech and Language Processing, September 23, 2018.

[7] A. Aizawa.(2003) .An information-theoretic perspective of tf–idf measures. Information Processing and Management 39 (2003) 45–65.

[8] How to Prepare Text Data for Machine Learning with scikit-learn (2003). https://machinelearningmastery.com/prepare-text-data-machine-learning-scikit-learn.

[9] http://nlp.stanford.edu/IR-book/html/html edition / naive-bayes-text-classification-1.html.

[10] S.Shwartz, S.and S. Ben-David. Understanding Machine Learning. Cambridge University Press, 2014.

ISBN 978-981-14-4787-7
Proceedings of 2020 the 10th International Workshop on Computer Science and Engineering (WCSE 2020)
Yangon (Rangoon), Myanmar (Burma), February 26- February 28, 2020, pp. 6-10

A Novel Clustering-based Class-asociation Rule Mining Method for Handling Class-Imbalanced Datasets

Tien-Dung Phan[1], Thanh-Tho Quan[2+] and Thi-Kim-Anh Vo[3]

[1]Faculty of Foreign Languages and Information Technology, People's Police College II, Vietnam
[2]Faculty of Computer Science and Engineering, Ho Chi Minh City University of Technology, Vietnam
[3]Faculty of Information Technology, Ho Chi Minh City Open University, Vietnam

Abstract. Class-association Rules (CARs) mining is a knowledge discovery technique with many practical applications. One of the extensions of mining CARs algorithm is to combine information about data classes to derive rules between item and class. However, in the class-imbalance field, it is difficult to mine the rules related to minor classes. One of the solutions is at first to cluster with the combination with CARs mining, then the items of minor classes can be grouped to some clusters. Thus, the corresponding rules will be easier to detect. The k-means clustering method is often used due to its fast computing speed. However, the clustering results of k-means are non-deterministic, so it may affect the clustering quality. In this study, we propose a new direction for combining k-means and Hierarchical Agglomerative Clustering, and continue with class-based association rule mining. Our method has the same execution time as the k-means method but has better clustering quality, so the generated rules are also more accurate, as illustrated in the experimental results.

Keywords: accuracy, CARs, classification, class-imbalanced dataset, clustering

1. Introduction

The classification method based on association rules has been researched and proved to be better than traditional rule-based methods such as ILA, ID3, etc. [1-3]. In fact, class-imbalanced datasets are quite common. This means there will be some layers with a number of samples that are superior to the others, which greatly affects the training process to classify and to predict classes. Especially, when classifying a classifier, if we choose an inappropriate minimum support threshold *(minSup),* the samples of minor classes will be unfrequent or the rules will be mined mainly the majority class samples.

For the above issues, Nguyen et al. (2016) proposed a clustering method using k-means algorithm to balance the number of samples of each class, then use CAR-Miner algorithm to mine the classification rules [4]. The study has demonstrated a significant improvement in the accuracy of comparisons between with and without implementation of class equilibrium. However, we realize that with this study, there are still limitations of k-means clustering technique and CAR-Miner algorithm:

- *k*-means clustering: although the execution time is relatively fast, it does not guarantee the similarity between the components in the cluster is good enough and it can not handle noises and outliers.

- CAR-Miner: although it is an efficient algorithm to mine classification rules based on MECR tree structure (Modified *Equivalence Class-Rules* tree). However, this algorithm consumes a lot of memory for storing the Obidsets (set of object identifiers containing itemset) of the itemset and requires computation time for the intersection of Obidset sets to each other. So, for a large database, this issue will become significant.

[+] Corresponding author. Tel.: +84 919890203
E-mail address: qttho@cse.hcmut.edu.vn

Based on limitations of k-means clustering methods when balancing class samples, we propose a new method to increase the similarity of data after being clustered in order to increase the accuracy for class prediction. Besides, we also apply CAR-Miner-Diff algorithm to solve the disadvantages of CAR-Miner presented in [4].

2. Mining Class-association Rules

Mining classification rules based on association rules mining (*Class Association Rules - CARs*) is to find a subset of association rules contained in the database [5]. The goal of mining classification rules based on association rule mining is: (i) Mining CARs meeting minimum support threshold (*minSup*) and minimum confidence threshold (*minConf*) ; and (ii) Build classifiers from CARs.

CAR-Miner is an improved algorithm of ECR-CARM algorithm developed by Nguyen et al in 2013 [6]. CAR-Miner mines class association rules based on MECR-tree structure. The MECR-tree structure (Modification of Equivalence Class Rule tree) is an improved tree structure from the ECR-tree structure, each node in the tree contains only a set of itemset with the following information:

- *Obidset* : a set of task object identifiers that contains itemset.

- (c_1, c_2,..., c_k): a list of integers, where c_i is the number of records in *Obidset* belonging to class c_i.

- pos:positive integers stores the position of the class with the highest count, ie. $pos = argmax_{i\in[1,k]}\{c_i\}$.

To solve the limitation of CAR-Miner algorithm based on MECR-tree structure, CAR-Miner-Diff algorithm was born. The CAR-Miner algorithm consumes quite a lot of memory for *storing Obidsets* of itemset sets and requires computation time for the intersection of *Obidset* episodes , this time becomes significant when we consider in a large database . CAR-Miner-Diff is an improved algorithm of CAR-Miner algorithm developed by Nguyen et. al [7]. CAR-Miner-Diff instead of storing the intersection between the Obidset sets, it only stores the difference between those Obidset sets (called *Diffset*), this leads to memory and speed of mining the association rules based on the tree structure are improved .

3. Combination of Clustering Algorithms and Mining Class-Association Rules

3.1. The Balance of the Class and the Clustering Algorithm Combination

In the data of class imbalance, the fact that some classes are in the majority will have a significant influence on the rule-based prediction process due to difficulties in selecting the minimum support thresholds. If the selected threshold is too high, leading to classes containing small samples would not be frequent, so there are no rules containing this class. If we select low threshold to mine the rules containing minority classes, the number of rules of the majority classes is still overwhelming so it also affects the class prediction stage. Therefore, we will balance the data of each class first, then perform the CARs mining.

In this paper, we use the concept of *intra-cluster similarity* to measure similarity between elements in a cluster. If this value is larger, the elements in the cluster will have higher similarity, thus clustering quality is better. Let *C* be a cluster with m elements, the similarity in cluster of *C*, denoted by Θ(C)) is the average of similarity between samples in *C* and is calculated by the following formula:

$$\Theta(C) = (\sum_{i=1}^{m-1}\sum_{j=i+1}^{m} Sim(W_i, W_j)) / (\sum_{i=1}^{m-1} i) \quad (1)$$

where *Sim* (W_i, W_j) is the similarity between the 2 samples in a cluster.

k-means is a simple clustering algorithm, the execution time is quite fast with the algorithm complexity is *O (nkd)* where *k* is the number of clusters, *n* is the number of samples and *d* is the number of times p. However, k-means does not guarantee the similarity in clusters is good enough. In contrast, HAC clustering algorithm has a longer execution time than k-means due to $O(n^2)$ complexity, but it returns cluster results with very similar clustering results [8]. Because we use the k-means method in the first step, the input of the HAC algorithm will be relatively small clusters, which makes the HAC algorithm run much faster in the second step. In addition, clustering can be controlled by HAC for better cluster quality.

3.2. K-means_Car-Miner-Diff

K-means algorithm is combined with the Car-Miner-Diff algorithm shown in Figure 1:

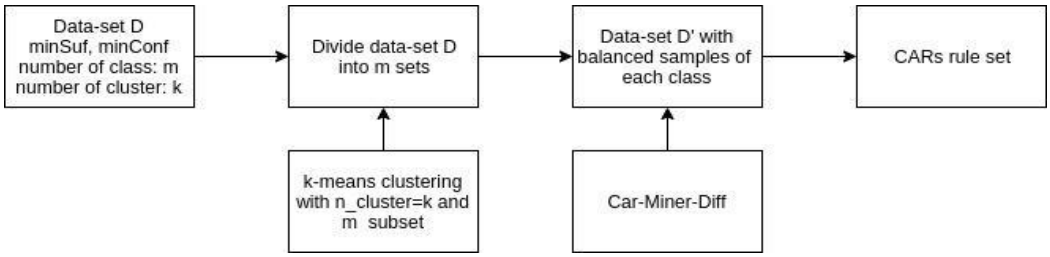


Fig. 1: Steps to combine the k-means algorithm with Car-Miner-Diff

INPUT: Dataset D, minSup, minConf, number of class m, number of cluster k.

OUTPUT: CARs rule set satisfies minSup and minconf.

3.3. K-means_HAC_Car-Miner-Diff

The k-means + HAC algorithm is combined with the Car-Miner-Diff algorithm shown in Figure 2:

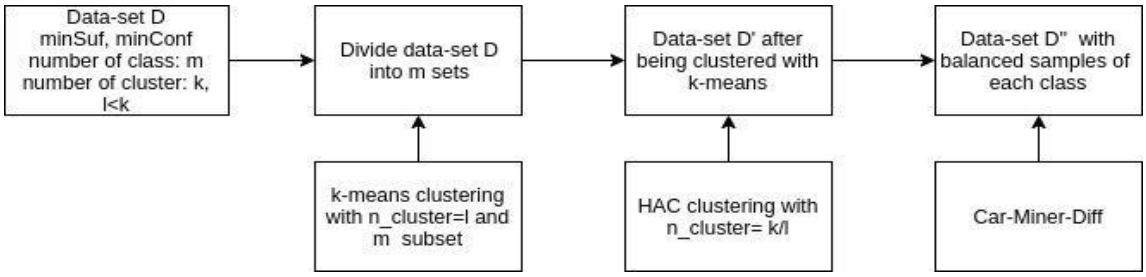


Fig. 2: Steps combination algorithm k-means + HAC with Car-Miner-Diff

INPUT: Dataset D, minSup, minConf, number of class m, number of cluster k, l (l<k).

OUTPUT: CARs rule set satisfies minSup and minconf.

First, we divide dataset D into m subsets corresponding to m values of class attribute. Let k be the number of data rows of the sub-data with the least number of samples.

Then, for each subset with the number of data rows that are greater than k, we apply k-means algorithm on that subset with l clusters (l<k). After that, we have a dataset D'.

For each small cluster which is the result of *k*-means clustering, we continue to apply HAC clustering with k/l clusters. With each cluster created, we only select a representative sample (the sample is the most similar to the center of the cluster). Thus, the result of each original subset will retain k samples and we have dataset D'' with balanced samples of each class.

Finally, we apply the Car-Miner-Diff algorithm on dataset D'' to mine CARs rule set.

4. Experimental Results

The standard empirical databases are taken from the UCI website http://mlearn.ics.uci.edu (Table 1)

Table 1: Experimental standard database

DATA SET	NUMBER OF PROPERTIES	NUMBER OF CLASSES	MODEL NUMBER	DESCRIPTION
Breast Cancer	9	2	683	- Class 0: 444 (65%)- Class 1: 239 (35%)
Chess	10	2	1200	- Class 0: 900 (75%) - Class 1: 300 (25%)
Diabetes	8	2	1400	- Class 0: 942 (67.3%) - Class 1: 458 (32.7%)
Tic-tac-toe	9	2	958	- Class 0: 332 (34.6%)- Class 1: 626 (65.4%)

4.1. Comparative Results on Accuracy

To compare and evaluate the results of accuracy of 03 algorithms: Car-Miner-Diff, k-means_Car-Miner-Diff, k-means_HAC_Car-Miner-Diff, the article uses 04 standard databases, minConf = 60% To proceed with the installation: Experimental results for the accuracy of the three algorithms are presented in Table 2

Table 2: Experimental results on the accuracy of standard databases (%)

Breast cancer	**MINSUP**	**0.5**	**0.3**	**0.1**	**0.08**
	k-means_HAC_Car-Miner-Diff	48.8966	63.1724	94.2759	95.1724
	k-means_Car-Miner-Diff	51.3235	61.5441	90.8088	92.6471
	Car-Miner-Diff	69.8049	69.561	79.3659	91.2195
Chess	**MINSUP**	**0.5**	**0.3**	**0.1**	**0.08**
	k-means_HAC_Car-Miner-Diff	51.1111	77.2222	83.8889	85.5556
	k-means_Car-Miner-Diff	78.9474	78.9474	78.9474	78.9474
	Car-Miner-Diff	75.5556	75.5556	75.5556	75.5556
Diabetes	**MINSUP**	**0.1**	**0.05**	**0.01**	**0.005**
	k-means_HAC_Car-Miner-Diff	75.6345	82.7411	86.802	87.3096
	k-means_Car-Miner-Diff	78.0749	78.0749	83.9572	84.492
	Car-Miner-Diff	71.9048	76.6667	79.7619	80.7143
Tic-tac-toe	**MINSUP**	**0.3**	**0.1**	**0.08**	**0.05**
	k-means_HAC_Car-Miner-Diff	53.2843	62.9902	74.0196	99.0196
	k-means_Car-Miner-Diff	50.7	70	69	88.5
	Car-Miner-Diff	68.4722	67.3611	67.3611	75.6944

The results from Table 2 show that for unbalanced class databases, the improved k-means_HAC_Car-Miner-Diff method results in better accuracy, especially for small minSup thresholds.

4.2. Comparison on Algorithm Execution Time

To compare and evaluate the results of the time of mining the rules of 02 algorithms: Car-Miner-Diff, k-means_HAC_Car-Miner-Diff, the article uses 04 standard databases minConf = 60% to proceed with the installation. Experimental results Perform the execution time between two algorithms are presented in Table 3.

Table 3: Experimental results on the implementation time on the standard database (ms)

Breast cancer	**MINSUP**	**0.5**	**0.3**	**0.1**	**0.08**
	Kmeans HAC Car-Miner-Diff	0	3	43	63
	Car-Miner-Diff	7	16	91	108
Chess	**MINSUP**	**0.5**	**0.3**	**0.1**	**0.08**
	Kmeans HAC Car-Miner-Diff	1	41	270	330
	Car-Miner-Diff	4	43	330	400
Diabetes	**MINSUP**	**0.1**	**0.05**	**0.01**	**0.005**
	Kmeans HAC Car-Miner-Diff	55	130	570	900
	Car-Miner-Diff	83	180	580	950
Tic-tac-toe	**MINSUP**	**0.3**	**0.1**	**0.08**	**0.05**
	Kmeans HAC Car-Miner-Diff	0	67	107	210
	Car-Miner-Diff	1	97	120	260

The results from Table 3 show that for databases with class imbalance, the k-means_HAC_Car-Miner-Diff improvement method due to processing with fewer samples after clustering should result in better processing time, especially for small minSup thresholds.

5. Summary and Future Work

In this paper, we have proposed an improved method combining k-means and HAC clustering techniques, implemented algorithms on standard databases, documented accuracy and execution time between subject methods, exported and original CAR-Miner-Diff algorithm for verification. At the same time, the paper also compares the similarity after clustering for 02 methods k-means and k-means_HAC.

In the future, we will continue to experiment on more types of databases with the increased number of classes to evaluate the applicability of the proposed improvement method. Furthermore, we will also apply this method to other types of classification such as decision trees, ILA, neural networks, etc.

6. Acknowledgements

This research is funded by Vietnam National University Ho Chi Minh City (VNU-HCM) under Grant B2018-20-07

7. References

[1] A. Veloso, W. Meira Jr., M.J. Zaki (2006). Lazy associative classification. In: Proc. of The 2006 IEEE International Conference on Data Mining (ICDM'06), Hong Kong, China, pp. 645- 654.

[2] A. Veloso, W. Meira Jr., M. Goncalves, H.M. Almeida, M.J. Zaki (2007). Multi-label lazy associative classification. In: Proc. of The 11th European Conference on Principles of Data Mining and Knowledge Discovery, Warsaw, Poland, pp. 605-612.

[3] A. Veloso, W. Meira Jr., M. Goncalves, H.M. Almeida, M.J. Zaki (2011). Calibrated lazy associative classification. Information Sciences 181(13), pp. 2656-2670.

[4] L.T.T. Nguyen, TMT. Tran, CH. Giang (2016). Exploiting association clustering rules with class-imbalanced dataset. In Proceedings the 10th National Conference of Fundamental and Applied IT Research (FAIR).

[5] J. Han, M. Kamber, Data Mining: Concepts and Techniques, 3rd Edition. Morgan Kaufmann Publishers, 2011

[6] L.T.T. Nguyen, B. Vo, T.P. Hong, H.C. Thanh (2013). CAR-Miner: An efficient algorithm for mining class-association rules. Expert Systems with Applications, vol.40, no.6, pp. 2305-2311.

[7] L.T.T. Nguyen, Ngoc Thanh Nguyen (2015). CAR-Miner: An improved algorithm for mining class association rules using the difference of Obidsets. Expert Systems with Applications, vol.42, pp. 4361-4369.

[8] K.T.Huynh et al. (2017), "A quality-controlled logic-based clustering approach for web service composition and verification", International Journal of Web Information Systems, Vol. 13 Issue: 2, pp.173-198

ISBN 978-981-14-4787-7
Proceedings of 2020 the 10th International Workshop on Computer Science and Engineering (WCSE 2020)
Yangon (Rangoon), Myanmar (Burma), February 26- February 28, 2020, pp. 11-15

Neighbor Search with Hash Map Indexing Technique for Complex Networks

Wai Mar Hlaing[1+] and Myint Myint Sein[2]

[1, 2]Geographic Information System, University of Computer Studies, Yangon, Myanmar

Abstract. Neighbor Search with Hash Map Indexing Technique is used to get the high performance when the optimal path is searched in the complex networks. This system can also give advice the public bus passengers about the travel route depend on the travel time and cost. Moreover, the proposed technique is highly performance one if it compares about the response time of many other popular cited shortest path algorithms. Especially it contains two main parts for finding the optimal path, the first one is dividing the complex large tree into small sub-trees using divide and conquer at an optimal threshold value. The second one is using heuristic neighbor search instead of searching the heuristic values of all expanded nodes at current level. Heuristic neighbor search and hash-map indexing technique is used together to reduce the time complexity when the heuristic values are searched dynamically depend on the user query to reach the target. The proposed system is faster than the popular bi-directional heuristic search A* algorithm, previously proposed combined forward and backward heuristic search algorithm and modified heuristic search algorithm. Road network and bus network in Yangon Region is used as the case study for spatial database.

Keywords: Divider and Conquer, Heuristic Search A*, Combined Forward and Backward Heuristic Search, Modified Heuristic Search

1. Introduction

Transportation contains as a major sector in all countries because it is related all other important sectors such as Education, Health and Business. Especially, most of the developing countries depend on the public transportation system. Shortest path finding is a considerable factor as the challenge for geographical information systems because geographical network is very complex. Shortest path finding algorithms adopt Heuristic Search and Classical Algorithms to find the optimal travel route. However, the problem of time complexity is still facing when the network structure is complex and large. Popular heuristic search algorithm A* is applied to find the shortest path during the short time because it does not need to expand all nodes while traversing the network to reach the target. A* is the high efficiency algorithm when the time complexity is compared among other Classical shortest path algorithms such as mature Dijkstra. However, A* algorithm needs to calculate all nodes in the open list repeatedly. In addition, A* need to pre-store the heuristic values between all nodes of the whole network database. Therefore, pre-processing time may be very long. For these problems, W. M. Hlaing proposed Combined forward and backward heuristic search algorithm [1]. This algorithm does not need to pre-store the heuristic values. It needs to calculate only the heuristic values of the neighbour nodes depend on the dynamically user query. CFBHS algorithm reduces the calculation time about finding the heuristic values of unnecessary nodes to find the optimal route. This algorithm is faster than the popular A* algorithm.

Some short path algorithms work in two directions to get the optimal route. It meets the short path result at a same point in forward and backward directions. It reduces the time complexity by working two

[+] Corresponding author. Tel.: + 959799533461.
E-mail address: waimarhlaing86@gmail.com;waimarhlaing@ucsy.edu.mm

directions during the same time when the short result is searched but it still needs to traverse all nodes in the whole graph. F. Islam uses A* algorithm with bi-directional search to meet the best efficiency [2].

CFBHS algorithm is faster than the A* algorithm and bi-directional Dijkstra. However, when the heuristic search A* algorithm is used together with bi-directional search, it is faster than the previous proposed CFBHS algorithm but the optimal result cannot be searched at sometimes because the path does not meet at the same point when the graph is traversed using A* in bi-directional. W. M. Hlaing proposed modified heuristic search algorithm [3]. It is faster than the CFBHS algorithm and A* algorithm in bi-directional search by working in one direction for finding the optimal route. At sometimes, modified heuristic algorithm cannot find the shortest path because it removes unexpectly some necessary nodes while searching the shortest path. However, the time efficiency is so fast if it compares the other algorithms such as CFBHS, heuristic search A*. Our work modifies the previously proposed CFBHS algorithm. It is faster than the CFBHS algorithm and the accuracy is better than the previously proposed modified heuristic search algorithm. And the main contributions for this work include:

- The short path finding in each subtree of a large tree network structure uses the hash map indexing approach to be easy and fast for retrieving the dynamically heuristic values.
- To increase the efficiency by reducing the calculation of unnecessary nodes repeatedly.
- To reduce the time complexity by studying the relations among the subtrees.

Indexing technique is used to be easy and fast for retrieving the required information from big data [4]. Hash map indexing is applied together with many techniques and algorithms such as Hash Map Indexing based on online querying, Approximate Nearest Neighbour Search and Content based Image Retrieval [5].

The rest of this paper is organized as follows. The newly proposed shortest path approach (i.e., the CFBHS with Hash Map Indexing) will be described in Section 2. Section 3 describes the comparison of node reduction, performance and accuracy among of different algorithms using statistical analysis. Section 4 concludes the paper.

2. Proposed Neighbour Search with Hash Map Indexing Technique

Retrieving the required information from Big Data is necessary for every sector. To be fast and easy for extracting the useful information from Geospatial Big Data becomes a challenge in Nowadays. Our proposed work will modify the previously proposed CFBHS algorithm by combining the mature approach known Hash Map Indexing Technique.

2.1. Hash Map Indexing Technique

Most of the research works when the data size is huge, these use the indexing technique to get high efficiency for data retrieving. Otherwise, at sometimes it is used as a caching technique and the recently used data are stored in hash table instead of database, file and other storage areas. Because, the time complexity about data accessing of hash map indexing is constant time. In our system, hash table is used as a memory cache. Hash table contains two parts: keys and values as shown in Fig. 1. Current Node and Target Node is saved together as a key. The respective values of the keys are stored by finding the heuristic distance values between these two nodes. These heuristic values are searched dynamically depend on the user query. The proposed system uses the number of hash table depend on the maximum degree of the vertex in the graph. The system uses only one hash table for sharing the heuristic values in this table among of the subtrees. All subtrees store their estimated heuristic values in the same hash table. Other hash tables are used to store the node that has the minimum value among of the current neighbor nodes.

2.2. Overview of the Proposed System

The proposed system is an extension of the work that previously published in the Global Conference on Consumer Electronics and already accepted the Journal of Internet Technology. When the size of the geospatial data is large, finding the optimal path in the short time becomes a challenge. The modified heuristic search algorithm is based on the CFBHS algorithm [3]. However, it works the short path finding in one direction. At first, this modified heuristic search algorithm splits the large tree into small subtrees. Each small subtree finds the local optimal shortest paths until the number of predefined threshold value is reached.

When the short path is searched from each subtree, it expands the tree only the neighbor nodes of the current node as the partial paths for the next level. In this way, it traverses the tree to reach the target. Euclidean Distance method is used for finding the distance values between the two nodes. For this case, the heuristic values between the current node and target may be duplicated among of the subtrees.

Our system solved about these problems using hash map indexing technique. Using the hash map indexing technique, the proposed system can reduce the unnecessary calculation times among of the subtrees. Fig. 1 describes the overview of the proposed system. The detailed works of the proposed system have described in the Journal of Internet Technology. Firstly, the system constructs the spatial database into graph database using Haversine Distance Method. Then, the tree network is separated into small subtrees using degree-based threshold calculation method. This pre-processing step needs to find the suitable threshold value only for the first time if the same database is used.

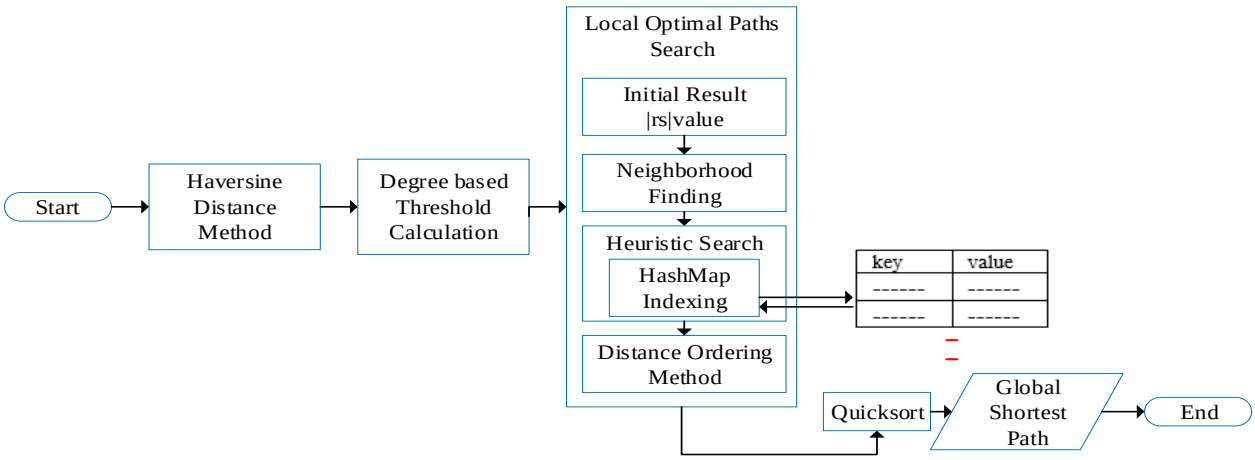


Fig. 1: System Flow

When the threshold value for dividing the tree has been defined, the system will define the number of results |rs| that will need to find from each subtree. When the shortest path is searched, the system considers the neighbor nodes of the current node to reach the target node. It does not expand all nodes because the time complexity may be high. In the previous proposed CFBHS algorithm, we need to find the paths in both forward and backward direction because it removed many nodes while going to target by considering only on the neighbors. However, it is faster than the popular heuristic A* algorithm. When the modified heuristic search is used, it is faster than CFBHS algorithm. This modified heuristic search algorithm's accuracy may be low because of finding the optimal paths in forward direction.

In this proposed system, we use hash map indexing technique when the heuristic value is searched to satisfy high accuracy and efficiency. The number of hash map table depends on the nature of the graph. If the maximum degree of a graph contains n, the number of hash map table also contains from Hash map table $H1, H2.., Hn$. The first table $H1$ is storing the heuristic values of expanded nodes for current query. The second table $H2$ is storing the comparison results of two nodes. For example, if node 'a' and node 'b' are compared, the system will store hash map key (a, b) and the node that possesses the smallest distance to reach the target among of these two nodes will be stored as the hash map value. Hash map table H3 will also store the comparison results among of three sub-nodes in current parent node. The remaining tables are saving the comparison results in this way. The heuristic estimated values are searched using straight line Euclidean distance. The smallest heuristic distance is choosing as the node that need for expanding next level. In this way, this partial path goes until the target node is reached. If the number of current results is less than the predefined result number |rs|, the system finds second local optimum path by removing, target node and a node before target occurs. After finding the local optimum results from each subtree, Quicksort algorithm is used to order the local optimal paths according to their total distance cost. Finally, the system will give the global shortest path.

2.3. Combining the Hash Map Indexing Technique in the Proposed System

When the current node of the tree is expanded into the neighbor nodes for the next level, the system must choose one node among of the neighbor nodes to go the target. The system needs to calculate the heuristic values between these neighbor nodes and the target node. The modified steps for the proposed system contain as follows:

1) Before calculating these heuristic values, the modified system will check the Hash Map table.
2) The system will check whether the current key value of the subtree already exists in the Hash Map Table. Key Name is the storing current node and target node together.

a. If the current key value already exists in the table, estimated heuristic value will not calculate again.
b. Otherwise, heuristic value will be calculated and stored together with key name in the Hash Map Table.

3) The system will compare the heuristic distance values among the expanded nodes of current level and target node.

a. Before comparison is made, the system will check whether this comparison already finish in their respective hash map table. If the number of expanded nodes of current level is four, hash map table H4 will be used. This table contains four nodes in a key. If the key already exists, the comparison result does not need to calculate again.
b. Otherwise, smallest distance values will be searched among of these nodes in current level of the tree.

Whenever the system decides to choose a node for the next level from each subtree, the modified steps that mentioned in above are used. When the network structure is complex and huge, these duplicate nodes calculation problems among of these subtrees are more evident. By combining the hash map indexing technique, previous nodes calculation problems among of the subtrees can be reduced.

3. Experimental Results

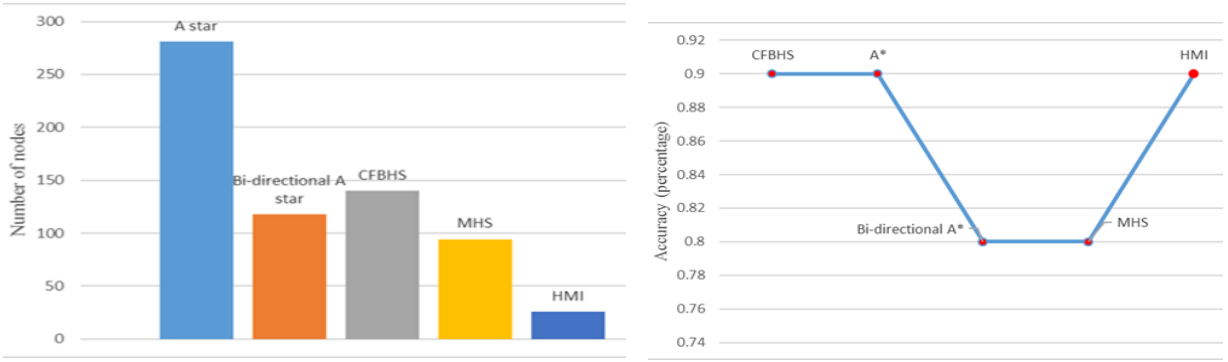
The spatial data of Yangon city in Myanmar are used for evaluating the proposed method. This section especially contains two main parts. The first part contains checking the number of compared nodes at each level of the network among of the algorithms using Yangon Downtown Network that contiains 44 nodes. It considers the junction points as the nodes of the Road network. Moreover, it compares the accuarcy among of different algorithms. The second part describes the statistical results about the accuracy of the algorithms such as A*, Bi-directional A*, Combined Forward and Backward Heuristic Search, Modified Heuristic Search and Hash Map Indexing with Heuristic Method using Yangon Bus Network. It is a complex hypergraph network and it contains about 4000 nodes and many edges.

3.1. Reducing Unnecessary Node Calculation and Accuracy Among of the Different Algorithms

Fig. 2 (a) describes the required nodes among of the short path finding algorithms that need to compare while finding the optimal route. Hash Map Indexing technique (HMI) need to compare only the heuristic values among of the few nodes. HMI is better than the previous proposed CFBHS and MHS algorithms according to the node reduction comparison. HMI calculates heuristic values dynamically depend on the user query and temporarily stored in the hash map tables. By adding hash map indexing technique in previous proposed CFBHS algorithm, it is significantly reducing the heuristic values calculation and nodes comparison jobs.

CFBHS algorithm is better than popular heuristic search calculation A* because CFBHS algorithm depends only on the neighbor nodes to choose a node for the next level. A* considers not only the expanded nodes of the current level but also all nodes in the open list. Modified Heuristic Search algorithm is better than Bi-directional A* algorithm because it works on the neighbor nodes in one directional search.

The accuracy of five short path finding algorithms using Yangon Downtown Network is compared in Fig. 2(b). The accuracy of these algorithms is searched using Eq.1. The percentage of accuracy is dividing the number of correct rules by the total number of rules. According to the comparisons, CFBHS, A*, CFBHS with hash map indexing is more accurate than the bi-directional A* and Modified Heuristic Search.



(a) (b)

Fig. 2: Number of comparison nodes and Accuracy among of Different Algorithms

$$\text{ACCURACY} = \frac{\text{CORRECTNESS}}{\text{TOTAL}} \times 100\% \quad (1)$$

3.2. Comparison About the Performance Among CFBHS, Modified Heuristic Search and Hash Map Indexing with Heuristic Method for the Yangon Region

Our system checks the accuracy and performance using different networks. Section 3.1 describes the performance and accuracy about Yangon Downtown Network. Section 3.2 in Current section also describes the comparison among of the algorithms using Complex Yangon Bus Network. HMI is the fastest heuristic search algorithm among of the other algorithms described in Fig. 3.

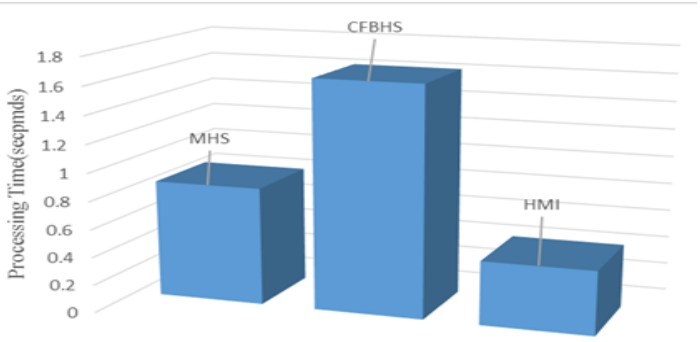


Fig. 3: The Processing Time among of Different Algorithms

4. Conclusion

Combined Forward and Backward heuristic search with hash map indexing technique is used to find the optimal route. This algorithm works only on the neighbor nodes while searching the optimal route. It divides the large tree into small subtrees. Therefore, it is faster than other popular heuristic search algorithms. This paper combines hash map indexing technique in CFBHS algorithm. This reduces unnecessary heuristic calculation and comparison time among of the subtrees. According to the statistical analysis and comparisons, by combining hash map indexing technique in combined forward and backward heuristic search algorithm, it outperforms to satisfy both conditions about accuracy and performance for different networks.

5. References

[1] W. M. Hlaing, S. Liu, and J. Pan. A Novel Solution for Simultaneously Finding the Shortest and Possible Paths in Complex Networks. J. Internet Technol., Sci. vol. 20, no. 6, pp. 1693–1708, 2019.

[2] F. Islam. A * -Connect: Bounded Suboptimal Bidirectional Heuristic Search. *Proc. of International Conference on Robotics and Automation (ICRA)*, pp. 2752–2758.

[3] W. M. Hlaing, M. M. Sein. K-means Nearest Point Search Algorithm and Heuristic Search for Transportation. 2018 IEEE 7th Glob. Conf. Consum. Electron., pp.779–780.

[4] J. Wang, W. Liu, S. Kumar, and S. Chang. Learning to Hash for Indexing Big Data - A Survey. 2015.

[5] P. Sadeghi-tehran. Scalable Database Indexing and Fast Image Retrieval Based on Deep Learning and Hierarchically Nested Structure Applied to Remote Sensing and Plant Biology, 2019.

ISBN 978-981-14-4787-7
Proceedings of 2020 the 10th International Workshop on Computer Science and Engineering (WCSE 2020)
Yangon (Rangoon), Myanmar (Burma), February 26- February 28, 2020, pp. 16-21

Analysis of Outlier Detection on Structured Data

Khin Myo Myat[1+] and Si Si Mar Win[1]

[1]University of Computer Studies –Mandalay, Myanmar

Abstract. Outlier detection has played an important role in all research areas for data analysis in various domains. Outlier is involved according to human error, sensors or mechanical faults and environment. It is detected to get data quality for all applications. On the other hand the good outliers can occur by chance in new contributions in research .The aim of this study is to discover new trend with outlier in dataset. The problem statement is how to detect the outlier analysis based on different datasets between before and after removing outlier. In this system, a box and whisker plot and the robust J48 algorithm are applied for outlier detection and classification.

Keywords: a box and whisker plot, classification, data Mining, J48, outlier detection

1. Introduction

Today the technology revolution has enabled us to gather massive amounts of data from different sources like social networks, sensor data, scientific data, biological data and networked systems data, etc. In real world data are incomplete, lacking attribute values and containing errors or outliers. So it will be important to have access to reliable data to make the decision and data preposition plays an important role in machine learning. Scientific experiments are especially sensitive situations when dealing with outliers. Outliers mean data points that are far from other data points. A data point is a discrete unit of information. Any single fact is a data point. The outlier can be a result of a mistake during data collection or it can be just an indication of deviation in the data. Outlier detection is an interesting problem in machine learning with the application from weather, computer security, financial and medical research domain.

In data mining, anomaly detection is the identification of unusual items, event or observations which raise uncertainty by differing obviously from the majority of the data. Unsupervised anomaly detection techniques detect anomalies in an unlabelled test dataset under the assumption that the majority of the instances in the dataset are normal by looking for instances that seem to fit least to the remainder of the dataset. In this paper, the different datasets and the WEKA toolkit were used to compares the accuracy which dealing with outliers and omitting outliers of data set. The box and whisker plot is used to outlier analysis on the different data set. It is needed around decisions why a specific data instance is or is not an outlier.

The paper is organized accordingly the introduction of outlier detection of data in section1. And then section2 describes the related works based on research paper. In section 3 methodologies is also described in section 4, the proposed system has been stated .In section 5, experimental results & analysis. And finally concludes with future work.

2. Related Work

In this section, the related work is presented in terms of outlier detection in preprocessing and classification of data mining. An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. In a sense, this definition leaves it up to the analyst to decide what will

[+] Corresponding author. Tel.: +95-9971323888
E-mail address: khinmyomyat558@gmail.com

be considered abnormal. Before abnormal observations can be signed out, it is necessary to characterize normal observations.

Classification has been successfully applied to a wide range of application areas, such as scientific experiments medical diagnosis, weather prediction, credit approval customer segmentation, target marketing and fraud detection [1, 2]. Decision tree classifiers are used extensively for diagnosis of breast tumor in ultrasonic images, ovarian cancer

Heart sound diagnosis and so on Arvind Sharma and P.C. Gupta discussed that data mining can contribute with important benefits to the blood bank sector. J48 algorithm and WEKA tool have been used for the complete research work. Classification rules performed well in the classification of blood donors, whose accuracy rate reached 89.9% [3].As medical records systems become more standardized and commonplace, data quantity increases with much of it going unanalyzed. Taking into account the prevalence of diabetes among men and women the study is aimed at finding out the characteristics that determine the presence of diabetes and to track the maximum number of men and women suffering from diabetes with 249 population using WEKA tool [4].

Data in most large databases is not perfect. The data called outliers or noise highlights a pattern that does not coincide with the pattern shown by the majority of data. So these data should be excluded from regular processing of data mining. Therefore, outlier detection (also known as anomaly detection) is required to find the outliers to improve the quality of data and to acquire a more accurate result of data mining. Outlier detection is an observation that deviates so much from other observations so as to arouse suspicions that it was generated by different mechanism. Outliers are probably generated because of measurement or executed error , etc., at the same time it could be considered as noise generally or decreased its effect in view of revised outlier value in order to pretreated data set.

R.Delshi Howsalya Devi et al.suggested a novel hybrid outlier detection based data mining algorithm to find the outliers in test data [5].

Data in most large databases is not perfect due to noise or outliers. . Removing outliers improves the quality of data and to acquire a more accurate result. Zheng et al. applied outlier detection in preprocessing step on the large cancer dataset. [6].

Some authors described the approach to detect anomaly without knowing the knowledge of anomaly class using an anomaly validation set for selecting hyper parameters of one class RBF-SVMs.They proposed for anomaly detection using classifiers to perform features to be a late fusion of hidden layer activation and residual error vectors and raw input signals.[7].

Recently, organizations have concluded that processing big data, especially the data coming from Twitter and Facebook can provide a significant impact on increasing the business's effectiveness and added values [8, 9].

Thomas et al. presented a numerical scheme for generating anomaly detection model that reduces to fitting distributions to data. They detected the problem of providing accurate ranking of disjoint time periods in raw IT system by monitoring data with their anomalousness. They suggested that their methods can be used to analyze various types of anomaly datasets. [10].

The data with outliers may also reduce clustering quality. So the author proposed a new outlier detection method to select initial cluster centers based on data density for adaptive K- Means Method [11].

The special characteristics of big data streams, such as transiency, uncertainty, multidimensionality, dynamic relationship, and dynamic data distribution, introduce new problems that make outlier detection for big data Streams more challenging [12]

3. Methodology

Data mining is a relatively a new technique to the world of information science. Detecting outliers, instance in database with unusual properties, is the data mining task. Now Outlier detection is the most researchable area in data mining field for knowledge discovery. An outlier is a rare chance of occurrence

within a given data set. There are good outliers that provide useful information that can lead to the discovery of new knowledge and bad outliers that include noisy data points.

The current section describes four main categories which include outlier detection in pre-processing, box and whisker plot, j48 algorithm for classification, Data Source and tools for experiments.

3.1. Outlier Detection

Outlier detection and analysis is an important data mining problem that goals to get anomaly points and behavior in data sets. It may be defined as the process of detecting and subsequently excluding outliers from a given dataset. To do so, outlier detection is vital to get the high quality of dataset. There are no standardized Outlier a branch of data mining has many applications in data analysis. An outlier may indicate bad data. For example, the data may have been coded incorrectly or an experiment may not have been run correctly.

Unlike many other methods of data display, they show outliers within a dataset. Outlier detection and analysis is an important data mining problem that goals to get anomaly points and behaviour in data sets. It may be defined as the process of detecting and subsequently excluding outliers from a given dataset.

3.2. Box and Wisher Plot

Box plots are useful for comparing datasets, especially when the datasets are large or when two or more data sets are being compared and when they have different numbers of data elements. It is a standardized way of displaying the distribution of data based on a five number summary. They are the minimum, first quartile Q1, median, third quartile Q3 and maximum. There are a few important vocabulary terms in a box and whisker plot method.

- Q1-quartile 1, the median of the lower half of the data set.
- Q2-quartile 2, the median of the entire data set.
- Q3-quartile 3, the median of the upper half of the data set.

IQR-Interquartile range, the difference from Q3 to Q1, which is the width of the box in the box and whisker plot. Extreme values- the smallest and largest values in a data set. In order to be an outlier, the data value must be larger than Q3 by at least1.5 times the interquartile range IQR or smaller than Q1 by at least 1.5 times the IQR when two or more data sets are being compared.

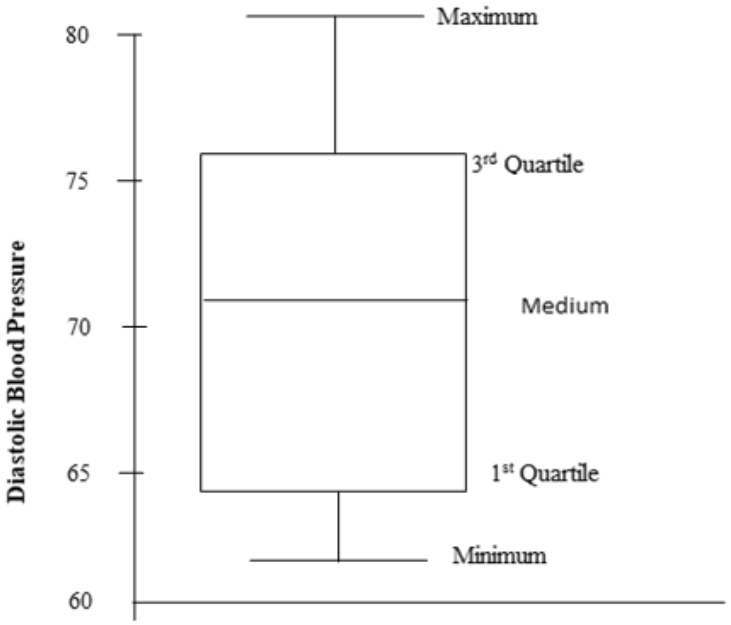


Fig. 1: Box and whisker plot

3.3. J48 Algorithm

J48 is an extension of ID3. The additional features of J48 are accounting for missing values, decision trees pruning, continuous attribute value ranges, derivation of rules, etc. In the WEKA data mining tool, J48

is an open source Java implementation of the C4.5 algorithm. The WEKA tool provides a number of options associated with tree pruning. In case of potential over fitting pruning can be used as a tool for pr écising. In other algorithms the classification is performed recursively till every single leaf is pure, that is the classification of the data should be as perfect as possible. This algorithm it generates the rules from which particular identity of that data is generated. The objective is progressively generalization of a decision tree until it gains equilibrium of flexibility and accuracy.

3.4. Data Source and Tool

We tested the variety of datasets on data mining tool such as WEKA. In this paper the WEKA toolkit were used to calculate the accuracy with outliers without outliers. WEKA is a collection of machine learning algorithms for data mining tasks. The algorithms can be used directly to the diabetes dataset. WEKA consists of tools for data preprocessing, classification, regression, clustering, association rules and visualization. It is open source software issued under the General public License. It is implemented in the java programming language and runs in any modern computing platform. It is portable and platform independent .it provides a graphical user interface for exploring and experimenting with machine learning algorithms on the datasets.

4. Proposed System

The proposed system with anomaly detection system detects outlier in the datasets by outlier detection method based on interquartile range and classification is done by using J48 classifier. In this paper, the comparison of accuracy which dealing with outliers and omitting outliers of dataset is presented. This paper discusses about the concept of outlier and outlier detection approaches. In this paper a combined approach such as a box and whisker for outlier analysis and J48 classifier for evaluating the effectiveness of outlier detection process. At first, a box and whisker plot algorithm is used for detecting outliers in the dataset. After the outliers have been removed, the data are given as input into a J48 classifier to evaluate the accuracy of dataset. Among all the classifiers, J48 is one of the best classifier in data mining.

By making a box and whisker plot Learn how to recognize potential outliers. Here, Diastolic blood pressure (mmHG) attribute in Diabetes dataset is considered for outlier detection. The values are 72.0, 66.0, 64.0, 66.0, 40.0, 74.0, 50.0, 500.0, 70.0, 96.0, 92.0, and 74.0

1. Arrange all data points from the lowest to highest
   40.0, 50.0, 64.0, 66.0, 66.0, 70.0, 72.0, 74.0, 74.0, 92.0, 96.0, 500.0
2. Calculate the median of the dataset.
   The median Q2= $\frac{70+72}{2}$ =71
3. Calculate the lower quartile Q1 is the data point below which 25% of the observations set.
   The lower quartile Q1=$\frac{64+66}{2}$ =65
4. Calculate the upper quartile Q3 is the data point above which 25% of the dataset.
   The upper quartile Q3=$\frac{74+92}{2}$ =83
5. Find the interquartile range
   Q3-Q1=83-65=18
   Lower limit = Q1-(1.5*IQR) =38
   Upper limit = Q3+ (1.5*IQR) =110

An outlier in example, it would have to be less than 38, which is the difference between Q1 (65) and IQR (18).Similarly, it would have to be more than 110, which is the adding Q3 (83) and IQR (18).

The values which are beyond these are extreme greater than 110 is 500 outlier.

The mean of our dataset with outliers is
(72.0+66.0+64.0+66.0+40.0+74.0+50.0+500.0+70.0+96.0+92.0+74.0)/12=179.66

The mean of our dataset without outliers is
(72.0+66.0+64.0+66.0+40.0+74.0+50.0+70.0+96.0+92.0+74.0)/11=69.45

Since the outlier can be attributed to human error and because it's inaccurate to say that the blood pressure was almost 180 mmHG, so should omit to this outlier

5. Experimental Result

Experiment are carried out on WEKA with 10 fold cross validation is where a given dataset is split into a 10 number of folds where each fold is used as a testing set at some point. It is used to perform statically analysis of the individual attributes in dataset. In this paper, the performance indicators such that RMSE, ROC, accuracy, Precision, Recall based on true positive and false positive are compared for the dataset using J48 algorithms and outlier detection algorithms. By comparing the accuracy and correctly classified attributes, suitable decision can be figure out.

The data sets within WEKA experiment are used to evaluate an attribute's efficiency by considering its mean, min, max, standard deviation and detect outliers. The outlier analysis is performed over all attributes in variety of data sets.

In this current research work, WEKA data mining tool is used to automatically detect outlier. We apply Filters option on unsupervised data and Inter Quartile Range (IQR) on a data set. After applying IQR two attributes are added ,outlier and extreme value .In this paper 7 experiments are performed ,in this some data sets does not have outliers means all instances are normal.

Table 1: Dataset with outlier and extreme

Dataset	Attribute before IQR	Instance	No. of Outlier	No. of Extreme
Diabetes	9	768	49	-
Glass	10	214	16	42
CPU	7	209	36	2
Segmant	20	150	114	424
German_Credit	21	1000	25	153
Weather	5	14	-	-
Iris	150	5	-	-

Table 2: Accuracy using j48 algorithm

Dataset	Accuracy (%)	Accuracy without outlier (%)	Accuracy without extreme (%)
Diabetes	73.82	74.40	74.40
Glass	66.82	68.68	69.93
CPU	96.65	98.48	100
Segmant	95.6	95.23	93.86
German_Credit	70.8	72.82	72.14
Weather	64.28	64.28	64.28
Iris	96.0	96.0	96.0

According to the result from experiment, the accuracies of classifier using dataset with without outliers increase in most datasets. Removing extreme also increases the accuracy in three datasets such as CPU, Glass and Diabetes. The accuracies using other datasets such as Weather, Iris and Segment are not different between before removing outlier and after removing outlier using WEKA tool. So outlier is not excluded in these datasets and it is retaining because it is considered as new trend for further work in data science.

6. Conclusions

The main goal of an outlier detection system is to detect the noisy or abnormal data in the dataset. Furthermore, it is equally important to detect errors at a preprocessing stage in order to reduce their impacts on further classification. This paper presents an outlier detection approach called Box and wisher plot where, the outlier dataset is measured by the J48 algorithm. The trained model consists of seven different datasets. Generally, removing outliers can provide the effective and efficient classification system. However, the experiments are especially sensitive situations when dealing with outliers, omitting an outlier in data can mean omitting information that denotes some new trend or discovery. In this situation, we should not omit outlier, assuming it is not due to an error, it represents a significant success in new discovery. So the outlier should be retained for further new trend research.

7. Acknowledgment

I would like to express my sincere gratitude to Prof. Dr Si Si Mar Win and Prof. Dr Kay Thi Win from the faculty of Computer Studies Mandalay. I also thanks University of Waikato for WEKA tool availability as an open source.

8. References

[1] Klemetinen, M., Mannila, H., Ronkainen, P., Toivonen, H., and Verkamo, A. I (1994) "Finding interesting rules from large sets of discovered association rules," CIKM.

[2] Tsumoto S., (1997)"Automated Discovery of PlausibleRules Based on Rough Sets and Rough Inclusion," Proceedings of the Third Pacific-Asia Conference(PAKDD), Beijing, China, pp 210-219.

[3] Arvind Sharma and P.C. Gupta ―Predicting the Number of Blood Donors through their Age and Blood Group by using Data Mining Tool‖ International Journal of Communication and Computer Technologies Volume 01 - No.6, Issue: 02September 2012.

[4] P.Yasodha, M. Kannan, "Analysis of a Population of Diabetic Patients Databases in WEKA Tool". International Journal of Scientific & Engineering Research Volume 2, Issue 5, May-2011 ISSN 2229-5518 Analysis of a Population of Diabetic Patients Databases in WEKA Tool

[5] R. Delshi Howsalya Devi, M.Indra Devi,A Novel Hybrid Algorithm for Outlier Detection Using WEKA Interface.International Journal of Applied Engineering Research, ISSN 0973-4562 Vol. 10 No.55 (2015)

[6] Zheng, B., Yoon, S. W., & Lam, S. S. (2014). Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. Expert Systems with Applications, 41, 1476–1482

[7] Jerone T. A. Andrews, Edward J. Morton, and Lewis D. Griffin, "Detecting Anomalous Data Using Auto-Encoders," International Journal of Machine Learning and Computing vol.6, no. 1, pp. 21-26, 2016

[8] B. Mantha,"Five Guiding Principles for Realizing the Promise of Big Data," *Business Intelligence Journal*, 2014. [Online] . Available: http://connection.ebscohost.com/c/articles/95066192/five-guiding-principles-realizing-promise-big-data. [Accessed : 04-Mar-2019].

[9] A. Fabijan, H. H. Olsson, and J. Bosch, "Customer Feedback and Data Collection Techniques in Software R&D: A Literature Review," Springer, Cham, 2015, pp. 139–153.

[10] Thomas J. Veasey and Stephen J. Dodson "Anomaly Detection in Application Performance Monitoring Data" *International Journal of Machine Learning and Computing, Vol. 4, No. 2, April 2014*

[11] Sarunya Kanjanawattana, "A Novel Outlier Detection Applied to an Adaptive K-Means," International Journal of Machine Learning and Computing vol. 9, no. 5, pp. 569-574, 2019

[12] Yogitaa, DT. A framework for outlier detection in evolving data streams by weighting attributes in clustering. *2nd IntConf Commun Comput Secur*. 2012;214–222

ISBN 978-981-14-4787-7
Proceedings of 2020 the 10th International Workshop on Computer Science and Engineering (WCSE 2020)
Yangon (Rangoon), Myanmar (Burma), February 26- February 28, 2020, pp. 22-26

Real-time Big Data Analytics for Feature Selection on Apache Spark

Lwin May Thant[+], Sabai Phyu

University of Computer Studies, Yangon, Myanmar

Abstract. Real-time data analysis is a key research in many domains. It can be applied to pre-existing or prescriptive models. The effective result is that monitor the account and review on a real-time action. Apache spark machine learning library Mllib can be distinct display place for real-time assessment foe extracting, transforming and selecting features and classification, clustering and frequent pattern mining. Feature selection is the detection in a group of feature what are the most relevant and removing the redundant data. Specifically, we made using the Apache spark tool and analyze the streaming time-series data using Mllib to extract the high qualitative feature in efficiently to get qualitative and high performance model.

Keywords: feature selection, apache spark, filter method, real-time data

1. Introduction

Nowadays, text data processing on large-scale is quickly important for many research area and business domains for real-time analytics. Data discovery and analytics using the model are basically in machine learning with real-time data. In an analysis on historical data is big-time to build the machine learning mode. In analytics phase, predict the model on live events. Apache spark platform are implemented several models for parallel and distributed processing on multiple machines. Moreover, spark Mllib is the implementation of machine learning framework to the distributed memory-based Spark architecture. It is platform independent and open-source libraries for big data implementation for distributed architecture and automatic data parallelization. Mllib can work a variety of machine learning functionalists such as extracting, transforming and selecting the features and classification, clustering and frequent pattern mining.

In those function, we implemented to extracting, transforming and selecting features on real-time data. This is the reason to reduce the computational cost of modeling and to improve the performance of the model. In this paper, three filters methods are used on high-dimensional classification with real-time data sets. We search the high accuracy methods with low run time. The remainder of this paper is organized as follows: Section 2, explain the related works. In section 3, we discuss the background knowledge for this paper and the three filter methods. Section 4, describe the real time streaming framework. Section 5 explain the experiments to compare the filter methods and analyze the results. In section 6, conclusion of our work.

2. Related Work

Many faster filter methods based on information theory, especially mutual information and SVM feature weights to mathematically evaluate the relevance and redundancy of data, optimizing their implementation through efficient parallelization is also crucial for challenge ultrahigh dimensional issued in big data [1, 2]. Author in [3], Evolutionary computation work with a filter feature selection algorithm. To obtain subsets of features from big data use the MapReduce archetype. To break down the original datasets into blocks of instances and learn from them a final vector of feature weights. The algorithm is implemented on the Spark framework and experiment show that increasing classification accuracy and runtime with big data. In [4],

[+] Corresponding author. Tel.: +09783788109
E-mail address: lwinmaythant@ucsy.edu.mm

network traffic feature selection on Spark with FSMS method. It is based on Fisher score and employ for subsets with a sequential featured search. On the Spark framework, this method decrease the classification and modelling time.

The work in [5], proposes the use case of X2 feature selection which is very popular in supervised learning pipeline. It is implemented the algorithm of the Scikit-learn Python machine learning library. The experiments run over the Data bricks platform and show that partitioning scheme of the data. Most of the feature selection algorithms depend on the number of features or instances, ReliefF depends linearly on both of them [6]. In [7] introduce the new general definition of L1-norm SVM (GL1-SVM) for feature selection and prove that solving the new proposed optimization problem reduces error penalty and enlarges the margin between two support vector hyper-planes. When facing with high dimensional instances, it can perform exactly well. In this paper, we focus the three methods for feature selection with real-time data on Spark framework. We evaluate the performance of those methods with respect to the classification accuracy and run time.

3. Background Theory

3.1. Feature Selection

Some of feature selection methods describe in this section. In filtering of feature selection methods, we will consider two kinds of methods upon a variable types: numerical and categorical and also two main groups of variables: input and output. In feature selection, input variables are typically provided as input to a model and indicates a classification predictive modelling problem. The statistical for filter-based feature selection are performed one input variable at a time with the target variable. These stasticial measures that any interaction between input variables is not considered in the filter process. In figure1, we demonstrate the feature selection methods based on input variable.

In feature selection methods can be classified into three categories according to their relationship with the classification algorithms: Filter, Wrapper and Embedded [8, 9]. In the wrapper method, the “usefulness” of a subset of a features is evaluated on the basis of the classifier performance [8, 9]. Embedded method exploit intrinsic characteristics of a given model to guide the feature selection process and choose feature which best contribute to the accuracy performance of the model [8, 9].

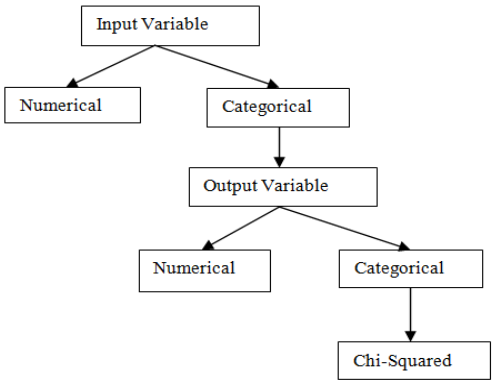


Fig. 1: Categorical flow of feature selection methods based on input variable

3.2. Feature Selection on Big Data

For a real time data sets, data collection, maintain, transmitting and processing within time interval are very difficult. So, very optimal technologies are required for processing the data without any fault and small amount of time interval. In big data research area, extracting require information from collections of data is one of the important fact. Moreover, efficient algorithms are need to apply to extract information in vast amount of data. In those day, classical big data analytics are special important problem to explore the data. Gartner [10] referred to big data in terms of volume, velocity, and variety, that is, the 3Vs, to which a further 2Vs were subsequently added, namely, veracity and value. Data analytics are interested in big data area, the large amount of instances while focusing to the feature aspect. To get more effective learning and prediction model, feature selection methods are needed for big data processing.

Removing of irrelevant, redundant and noisy features from complex features is the key research area in big data. For large-scale data, new feature selection techniques are very essential to obtain optimal information. We will discuss the feature selection framework for big data, this is the main challenges made to develop the classical approach to the new big data research trend. We also analyze the complexity arising from parallelization of the operations. We will consider the following phase while maintaining certain features.

Phase 1: Columns Transformation - The access pattern selection is variety, which operate on rows or columns. Although this may be considered, it can be decrease performance when compute relevance and redundancy in feature selection method. For distributed framework like Spark, this issue is conspicuously important when the data partition is a big effect on performance.

Phase 2: Broadcasting - Be minimal to avoid CPU usage, all features values have been grouped and partitioned into different partitions. By replicating the output feature and the final selected feature, data shift is reduced in each iteration.

Phase 3: Precomputed Data Caching - Subsequent marginal and join proportions are also performed. So, redundancy computation by features are reduced.

Phase 4: Greedy Approach - This phase reduce the common of complexity in feature selection by selecting the number of features. Greedy search is selected only one feature in each iteration.

3.3. Apache Spark on Big Data

Apache Spark framework for big data analytics is in-memory programming model and libraries for scalable machine learning, graph analysis, data streaming and structured and unstructured data processing. It is a cluster computing framework and open source project, with an increasing development in both academia and industry especially who are beginners in this area.

Extraction of the right features is one of the most challenging tasks in big data processing. Although solution of this task is Spark Mllib provide different methods for feature selection. While processing of feature extraction is to extract features from raw data feature transformers can be used for scaling, normalization, converting features, modifying features and so on. The machine learning library of Apache Spark contains methods for selecting subsets of features from larger sets of features In figure 2 show the framework for feature selection with Spark in big data.

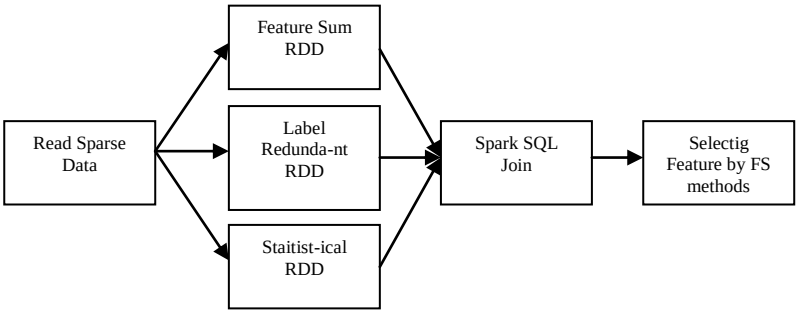


Fig. 2: Feature Selection Framework with Spark

3.4. Featuring Selection Methods

3.4.1. Kruskal-test

Filter Kruskal-test is a non-parametric method which checks whether the samples originate from the same distribution or not. In the first step, the features are sorted in ascending order and then rank the sorted data point and if any tied values present then give an average rank. Finally, calculate the filter score for feature X_k.

3.4.2. Chi-squared Test

Chi-squared test X^2 is the prominent statistical test and perform independent between the observed and expected distribution of a feature. The chi-squared use the value of X^2 as score. The value of X^2 statistical is directly proportional to the dependency between the class variable and feature.

3.4.3. Relief

Relief is the analysis of the quality of attribute in their values distinguish between instances. The Relief search the two nearest neighbors: one for within one class and another for different class. This method select randomly and regarded scored as x is the sum of weighted differences within same class or different class. In above mention methods are used in this paper and we will extract with optimal features with best method.

4. Real Time Streaming Big Data with Spark

The streaming data contains a wide variety of thousands of data sources. The real time streaming framework are working with these flow. In the data source layer, data are collecting from file system, cloud bucket databases and real time stream from IoT devices. Apache Nifi is a data collection tool that allows to send, receive, route, transform and sort in an automatic way in ingestion layer. In this layer, data are transformed by predefined processors. For ingestion of real time data, Apache Kafka is used and this is messaging system. Apache Kafka allow the partition to parallelize by splitting the data. Each partition are work on a separate machine in parallel. Apache Spark framework is the structured streaming model to provide fast, scalable, fault tolerant low latency. It is perform batch and streaming processing. It is provided in transformation layer.

When reading streaming data from Kafka through Apache Spark, Kafka send a data frame. Kafka cluster specify the location of data frame which are reading the data frame. In spark structure wait for 1 second and batches all the events during the time between micro batches. When finishing the micro batch processing, new batch is received and schedule again. However, latency does not decrease in streaming execution. Processing flow are shown in Figure3.

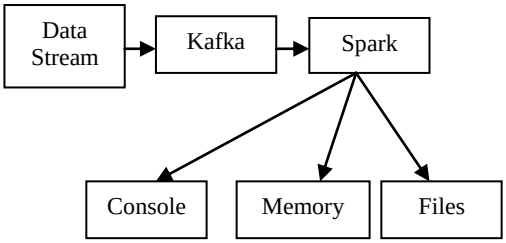


Fig. 3: Spark Flow for real time streaming

Table 1: Summary Dataset

Dataset	Size	Rows	Features
1	1.3 GB	4,203,855	11,556,704
2	2.67 GB	8,407,733	20,215,650

Table 2: Comparison of classification performance on Dataset 1

Comparison of feature selection methods	Support Vector Machine	K Nearest neighbor	Naïve Bayes	Random Forest
Kruskal-test	0.80	0.83	0.80	0.82
Chi-squared test	0.78	0.81	0.82	0.85
Relief	0.73	0.69	0.71	0.79

Table 3: Comparison of classification performance on Dataset 2

Comparison of feature selection methods	Support Vector Machine	K Nearest neighbor	Naïve Bayes	Random Forest
Kruskal-test	0.91	0.92	0.93	0.95
Chi-squared test	0.87	0.90	0.91	0.92
Relief	0.87	0.77	0.86	0.85

5. Experiment and Result

Evaluation of feature selection methods can be compared in many ways. We choose the way is to classify the performance of features by selected methods. When featuring the significant instance correctly with featuring methods, we will classify the features with perfect performance with small amount of features. If we will know the exactly features, we will calculate the feature selection methods effectively. But, real

time data are not static and thus we calculate the data with unknown various features. Since the exactly features are reported the real time data, it can be applied the different feature selection methods. Over a various separate set, we used the feature selection for training set and classification for validation set. So, featuring selection methods are used to obtain the valid feature set in training feature. Classification performance are calculated on the validation set for feature selection methods. To measure the classification performance, AUC values are used and increase the AUC values we will better classified.

In our implementation, we used the Apache Spark Mllib for each classification algorithms and summary of data set shown in Table 1. Table 2 shows the data set 1 of maximum value for the combination of feature selection methods and classification techniques. For dataset2, we compared the four classification performance for features. 10- Folds cross validation method are used for calculation of AUC values and 9 folds are feature selection and training. For the combination of selection methods and classification algorithms repeating the process 10 times over the ten distinct cases. Table 2 show the combination of selection methods and classification algorithms of maximum AUC values. We can be said what selection method is better performance in feature selection and which is low run time.

6. Conclusion

In this work, a filter-based feature selection method has been applied to real time streaming data. We calculated performance on two data sets: one from the news twitter and the other from Amazon. In this paper, on both data sets were compared with filter-based feature selection methods and classification performance in selecting features for different classification algorithms. We can trained a model with various feature selection methods and classifiers what methods are better performance on selecting features. Finally, in evaluating the AUC values we analyzed their performance. Thus we can conclude that Kruskal-test is a suitable technique for feature selection on streaming data.

7. References

[1] "A Large Scale Filter Method for Feature Selection Based on Spark", 2017 4th IEEE International Conference on Soft Computing and Machine Intelligence.

[2] Jazeera K.U. and Julie M. David. Issues, challenges, and solutions: big data mining. Sixth International Conference on Networks & Communications. DOI: 10.5121/csit.2014.41311.

[3] D Peralta, S Del R ó, S Ram írez-Gallego, I Triguero, Jose M. B F Herrera. Evolutionary Feature Selection for Big Data Classification: A MapReduce Approach. Mathematical Problems in Engineering Volume 2015 (2015)

[4] Yong Wang, Wenlong Ke, Xiaoling Tao.A Feature Selection Method for Large-Scale Network Traffic Classification Based on Spark. Information (2078-2489). 2016, Vol. 7 Issue 1, p1-11. 11p.

[5] M Nassar, H Sofa, A Al Mutawa, A Helal, Iskander GABA "Chi-squared Feature Selection over Apache Spark M Mandal, A Mukhopadhyay.

[6] An Improved Minimum Redundancy Maximum Relevance Approach for Feature Selection in Gene Expression Data. Jul 2016.

[7] Hai Thanh Nguyen, Katrin Franke, and Slobodan Petrovi'c, "On General Definition of L1-norm Support Vector Machines for Feature Selection," International Journal of Machine Learning and Computing vol. 1, no. 3, pp. 279 283, 2011.

[8] C Liu, W Wang, Q Zhao and M Konan. A new feature selection method based on a validity index of feature subset. Pattern Recognition Letters, Volume 92, 1 June 2017, Pages 1- 8.

[9] S Ramırez-Gallego, H Mourino-Talın, Francisco Herrera. An Information Theory Based Feature Selection Framework for Big Data under Apache Spark. Journal of latex class files, vol. 13, no. 9, September 2014.

[10] Y-W Chang, C-J Lin. Feature ranking using linear SVM. Proceedings of the Workshop on the Causation and Prediction Challenge at WCCI 2008, PMLR 3:53-64, 2008.

ISBN 978-981-14-4787-7
Proceedings of 2020 the 10th International Workshop on Computer Science and Engineering (WCSE 2020)
Yangon (Rangoon), Myanmar (Burma), February 26- February 28, 2020, pp. 27-35

Comparative Results of Dependent and Independent Variables Focused on Regression Analysis Using Test-Driven Development

Myint Myint Moe[1+], Khine Khine Oo[2]

[1]University of Computer Studies, Yangon, Myanmar
[2]University of Computer Studies, Yangon, Myanmar

Abstract. Test-Driven Development is a software engineering technique and should be tested previous the code, which make sure to read the unit test. The goal of this paper is to examine product quality and programmer productivity on the number of tests for the consequence of test-driven development. This system builds the acceptance test suite metric and the ordinary least squares method of regression analysis to assess the impact of the process on dependent variables and independent variables. The results of this paper are that if developer productivity is the actual effect, external code quality will be fewer decreased and external code quality is the actual effect if developer productivity will be fewer reduced. TDD responds to assist the delivery of high-quality products both operational and technical perspective while enhancing developers' productivity. TDD leads to less defects and fewer debugging period which actual code can be assured by writing tests first and thus serving the developer get a finer understanding of the software requirements. This proposed system evaluates the ordinary least squares and the acceptance test suite metric of regression analysis based on a fixed time-frame.

Keywords: Test-Driven development, Unit test, no: of tests, External Quality, Developer Productivity

1. Introduction

Test-driven development is the foundation of software development but it responds unit tests previous production code. Test-driven development is part of the agile code development approach and drive from Extreme Programming and the Agile Platform. Before the code development, developers encourage to compose tests [1- 2]. The possible of TDD describes various positive effects. TDD isn't a testing approach, yet rather a development and design method in which the tests are composed before the production code. During the implementation, the tests are added step by step and when the test is passed, the code is refactored to improve the inside structure of the code, without changing its outside behavior. TDD cycle is repeated until the whole functionality is implemented. For each little function of an application, TDD begins with designing and developing tests. First, the test is created that distinguishes and approves what the code will do in the TDD approach. Make the code and after that test in the typical testing process. The developer can be self- assurance that code refactoring is not destroyed any existing functionality for re-executing the test cases [3- 4].

This paper is structured as follows. The issue of Test-Driven Development initiated in Section (1). The obviousness of the number of tests, external code quality and, developer productivity on test-driven development (TDD) expressed in Section (2). A background of the whole procedure of test-driven development is presented in Section (3). Section (4) describes the contribution of the interrelation of the number of tests, external code quality, and developer productivity. Next, Section (5) shows the proposed system of this paper. Section (6) in this paper describes discussion and comparison of results. Finally, the conclusion of this paper describes in Section (7).

[+] Corresponding author. Tel.: +09448018175
E-mail address: myintmyintmoe.ucsy.1971@gmail.com.

2. Objectives

One of the approaches of software progression was test-driven development. In recent years, this approach has become familiar in the industry as a requirements specification method. TDD is intended to make the code clearer, simple and bug-free [4- 5]. The goal of the proposed system analyzes the consequence of dependent variables and independent variables on TDD**.** It observes the nature of the correlation between the number of tests (TEST) and external code quality (QLTY), and the correlation between the number of tests (TEST) and developers' productivity (PROD). The good points of Test-Driven Development upgrade software quality and accelerate the testing process. This approach more productive program code and make fewer efforts per line of code. By decreasing code complexity supporting, the proposed system validates the exactness of all codes and allows developers assurance. It is used persistently over time and motivates developers to create higher code quality. This system assures the correctness of the code and helps developers' gain a better understanding of the software requirements which leads to fewer defects and less debugging time. This intends more productive and make fewer efforts per line of code.

3. Background Theory

TDD builds up the early development of tests, at the time changes are welcomed and advanced with functional components. So, correcting defects is made earlier in the process. Test-Driven Development is a coding technique. Kent Beck (inventor of Extreme Programming and JUnit) invents Test-Driven Development refers to a style of programming where three activities are closely interlinked. Three activities are Coding, Testing, and Design. At first, its key idea is to execute initial unit tests for the code, must be implemented, and then implement the actual feature of it. One of the features of software system requirement is tackled subtask or user stories, which are designed to easily express and understand. These can be easy to change by the end-user as they like during the project's handle time [5-7].

3.1. Test-Driven Development Process

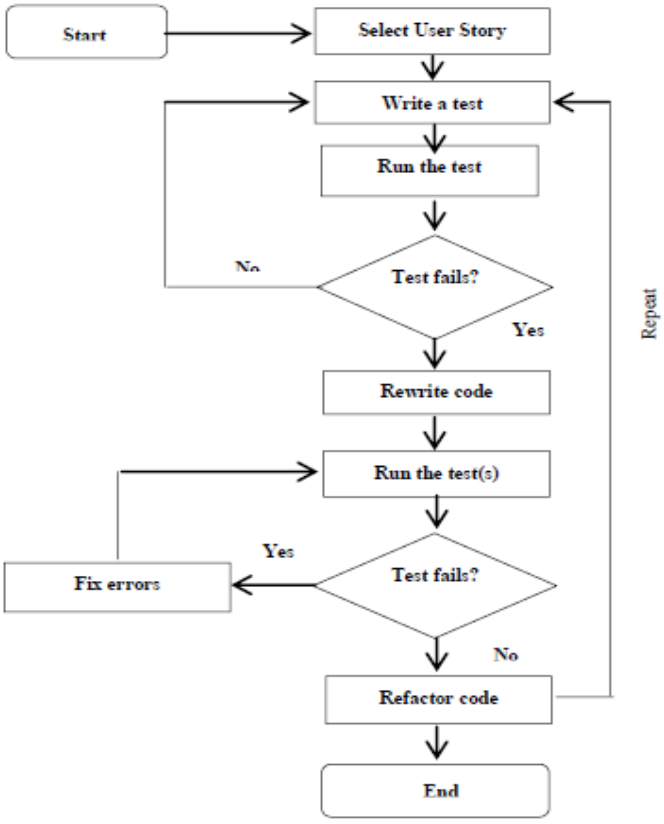


Fig. 1: Test-Driven Development flow

The TDD action is introduced in Figure 1, and includes the following steps:

(1) Select a user story,

(2) Write a test that fulfills a small task of the user story and run this test. Then produces a failed test,

(3) Re-write the production code necessary to implement the feature,

(4) Execute the pre-existing tests again, where any failed test is existed. When the code is true effectively and finally goes to the refactoring stage.

(5) When the refactoring stage is completed, the actual production code is manufactured and the user can select a new user story again. This method constructs some benefits, focus on the responsibility of increasing the quality of the software product and the productivity of programmers.

4. Contribution

The contribution of this paper observed the number of tests (TEST), external code quality (QLTY) and developer productivity (PROD). The number of tests is measured by the count of a single JUnit test case. External code quality is proposed the percentage of acceptance tests passed for the implemented user stories. The developer productivity is proposed the percentage of implemented user stories.

5. Proposed System

In this proposed system, the acceptance test suite metric and the ordinary least squares method of regression analysis apply to measure the number of tests, external code quality, and developer productivity.

5.1. Research Questions

This system concentrates to evaluate two outcomes on the following system: external code quality based on the number of tests and developer productivity based on the number of tests.

RQ1 (RQ-QLTY): Does a higher number of tests indicate higher quality?

RQ2 (RQ-PROD): Does a higher number of tests indicate higher developer productivity?

The notion of external code quality in RQ-QLTY and productivity in RQ-PROD are based on the acceptance test suite metric and the ordinary least squares method of regression analysis.

5.2. Method of Acceptance Test Suite Metric

In the proposed system, the acceptance test suite metric is used by analyzing to explore possible interactions such as number of tests, external code quality, and developer productivity. The acceptance test suite metric is a form of mathematical regression analysis [7-9]. Regression analysis is used to investigate the relationship between two or more variables and estimate one variable based on the others. Regression analysis is a powerful statistical method that allows for analyzing the relationship between two or more outcome variables of interest. QLTY and PROD are the dependent variables. TEST is the independent variable. QLTY defined as the percentage of acceptance tests passed for the implemented tackled tasks. PROD measured as the percentage of implemented tackled tasks. Table 1 provides the raw data used in the assessment. To calculate this low-level metric, this system uses an automated tool. The limited-time necessary to complete the task had an impact on the metric. In regression analysis, dependent variables are established on the vertical y-axis, while independent variables are established on the horizontal x-axis.

5.2.1. Number of Tests (TEST)

Number of Tests (TEST) is defined as numbers of JUnit assert statements within the unit test suite written by the participants while tackling the task. The numbers of test development as a single JUnit assert statements. TEST assessed by the count of the JUnit test cases. TEST is a ratio measure in the range [0, ∞]. The formula for calculating TEST is defined as [9]:

$$\text{TEST= no: of subtasks out of results from the no: of input subtasks} \quad (1)$$

$$\text{TEST = the numbers of JUnit assert statements within the unit test suite}$$

5.2.2. External Code Quality

The metric for external quality QLTY based on the number of tackled subtasks *(#TST*) for a given task. A subtask as tackled assesses if at least one assert statement in the acceptance test suite associated with that subtask passes. QLTY is a ratio measure in the range [0,100].

The number of tackled subtasks (*#TST*) is defined as:

$$\#TST = \sum_{i=0}^{n} \begin{cases} 1 & \#Assert_i(Pass) > 0 \\ 0 & otherwise \end{cases} \quad (2)$$

#TST = the number of tackled subtasks

n = the total number of subtasks

The formula for computing QLTY is defined as [10]:

$$QLTY = \frac{\sum_{i=0}^{\#TST} QLTY_i}{\#TST} \times 100 \quad (3)$$

$QLTY_i$ = the quality of the i^{th} tackled subtask

Where $QLTY_{i\ is}$ the quality of the i^{th} tackled subtask and $QLTY_i$ is defined as:

$$QLTY_i = \frac{\#Assert_i(Pass)}{\#Assert_i(All)} \quad (4)$$

#Assert$_i$ (Pass) = the number of JUnit assertions passing in the acceptance test suite associated with the i^{th} subtask

#Assert$_i$ (All) = the total number of JUnit assertions in the acceptance test suite associated with the i^{th} subtask

For example, supposing that a person assesses thirteen tackled subtasks (*#TST* = 13), this means that there are thirteen tackled subtasks for which at least one assert statement passes in the test suite. Let us assume that the acceptance test of the first analyzed tackled task contains 3 assertions, out of which three are passing. The acceptance tests of the fourth tackled task contain 10 assertions, of which three are passing and so on.

i.e. $(QLTY_4 = \frac{\#Assert_4(Pass)}{\#ASSERT_4(All)} = \frac{3}{10} = 0.3)$

$(QLTY = \frac{\sum_{i=1}^{\#TST} QLTYi}{\#TST} \times 100 = \frac{1+1+1+0.3+1+1+1+1+1+1+1+1+1}{13} \times 100 = 95)$

Table 1: Summary of acceptance tests used to calculate the metrics of Bowling Scorekeeper data-sets [10].

Task	Test	Assert
T1	3	3
T2	3	3
T3	2	2
T4	3	10
T5	5	5
T6	6	6
T7	8	8
T8	5	5
T9	5	5
T10	4	4
T11	2	2
T12	3	3
T13	2	2

Table 2: Solution of QLTY

Task	Test	Assert	QLTY
T1	3	3	1
T2	3	3	1
T3	2	2	1
T4	3	10	0.3
T5	5	5	1
T6	6	6	1
T7	8	8	1
T8	5	5	1
T9	5	5	1
T10	4	4	1
T11	2	2	1
T12	3	3	1
T13	2	2	1
	51	**58**	**95**

5.2.3. Productivity

The productivity metric (PROD) describes the amount of work successfully performed by the subjects. PROD is a ratio measure in the range [0,100]. The metric of PROD is calculated as follows [10]:

$$\text{PROD} = \frac{\#Assert(Pass)}{\#Assert(All)} \times 100 \quad (5)$$

For example, assume a person implements a tacked task with a total of 58 assert statements in a test suite. After running the acceptance test suite against the person's solution, 51 assert statements are passing.

i.e. $(\text{PROD} = \frac{\#Assert(Pass)}{\#Assert(All)} \times 100 = \frac{51}{58} \times 100 = 88)$

5.2.4. Assessment

The image below is a scatter plot**.** Scatter plots are used when this paper want to show the relationship between two variables. Scatter plots are called relationship plots because they show how two variables are interrelated. This analytical tool is most often used to show data correlation between two variables. This system expects that the regression analysis of the information compiled from the external code quality over the number of tests by TDD responds positively to questions RQ1. In the same way, this system expects that the regression analysis of the information compiled from the developer productivity over the number of tests by TDD responds slightly decrease to questions RQ2. In figure 2, the external code quality over the number of tests is improved by measuring the acceptance test suite metric of quality (QLTY). In figure 3, the developer's productivity over the number of tests is slightly decreased by measuring the acceptance test suite metric of productivity (PROD).

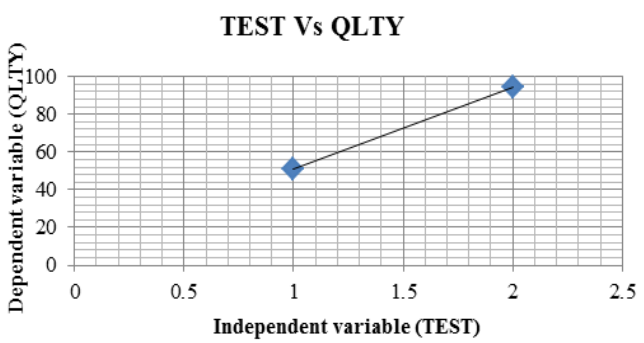


Fig. 2: QLTY is on the function of TEST

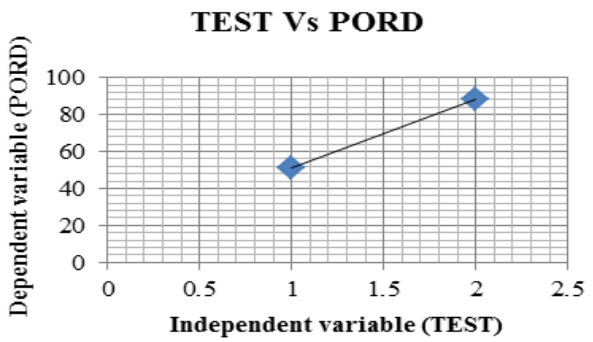


Fig. 3: PROD is on the function of TEST

5.3. Method of Ordinary Least Squares

QLTY and PROD are the dependent variables. TEST is the independent variable. The data set consisting of TEST, QLTY and PROD attributes were analyzed to discover outliers using both z-score and modified z-score methods [11-12]. Table 1 provides the raw data used in the assessment. In the proposed system, the ordinary least squares method is used by analyzing to explore possible interactions such as number of tests, external code quality, and developer productivity. TEST assessed by the count of the JUnit test cases. This is a ratio variable within the range [0, ∞]. QLTY explicated as the percentage of acceptance tests passed for the implemented stories. PROD measured as the percentage of implemented stories. The ordinary least squares method is a form of mathematical regression analysis used to determine the line of best- fit for data points. Each point of data describes the association between a known independent variable and an unknown dependent variable. In regression analysis, dependent variables are depicted on the vertical y-axis, while independent variables are depicted on the horizontal x-axis. The line of best-fit determined from the least-squares method has an equation that states the user story of the correlation between the data points. Line of best-fit equations may be determined by computer software models, which include a summary of outputs for analysis, where the coefficients and summary outputs explain the dependence of the variables being tested. The b_1 is the slope of the regression line. Thus this is the amount that the Y variable (dependent) will convert for each 1 unit convert in the X variable. The b_0 is the intercept of the regression line with the y-axis. Y-hat = $b_0 + b_1(x)$ is the illustrative regression line. This paper must determine b_0 and b_1 to draw this line. Y-hat is the predicted value of Y, and it can be obtained by plugging an individual value of x into the equation and calculating y-hat.

The ordinary least squares of regression analysis are computed as the formulas:

$$\text{Mean of TEST } (\bar{X}) = \frac{\Sigma X}{N} \quad (6)$$

ΣX = sum of all the individual #TEST data set
N = total number of # TEST data set

$$\text{Mean of QLTY (or) PROD } (\bar{Y}) = \frac{\Sigma Y}{N} \quad (7)$$

ΣY = sum of all the individual QLTY or PROD data set
N = total number of QLTY (or) PROD data set
Predicted value of Y,

$$b_0 = \hat{y} - b_1 x = \text{mean } (\bar{Y}) - b_1 * \text{mean } (\bar{X}) \quad (8)$$

b_0 = the intercept of the regression line with the y-axis

$$b_1 = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\Sigma(X - \bar{X})^2} \quad (9)$$

b_1= the slope of the regression line

$$\hat{y} = b_0 + b_1 x \quad (10)$$

$\hat{y}$= the sample regression line

The table-1 dataset consisting of TEST, QLTY and PROD attributes was tested to find outliers using both z-score and modified z-score method**s.** This paper used the dataset from [12].

Table 3: Dataset Used in the Assessment

TEST	QLTY	PROD
16	69	100
10	28	46
14	49	92
17	72	100
14	78	92
17	75	100
25	60	69
11	69	85
6	26	46
5	43	31
14	68	100
13	86	100
13	68	85
8	11	46
11	75	54
10	55	85

Table 4: Data of Computation for Correlation of TEST and QLTY

X (TEST)	Y (QLTY)	$X - \bar{X}$	$Y - \bar{Y}$	$(X - \bar{X})^2$	$(X - \bar{X})(Y - \bar{Y})$
16	69	3.25	10.69	10.56	34.73
10	28	-2.75	-30.31	7.56	83.36
14	49	1.25	-9.31	1.56	-11.64
17	72	4.25	13.69	18.06	58.17
14	78	1.25	19.69	1.56	24.61
17	75	4.25	16.69	18.06	70.92
25	60	12.25	1.69	150.06	20.67
11	69	-1.75	10.69	3.06	-18.70
6	26	-6.75	-32.31	45.56	218.11
5	43	-7.75	-15.31	60.06	118.67
14	68	1.25	9.69	1.56	12.11
13	86	0.25	27.69	0.06	6.92
13	69	0.25	10.69	0.06	2.67
8	11	-4.75	-47.31	22.56	224.73
11	75	-1.75	16.69	3.06	-29.20
10	55	-2.75	-3.31	7.56	9.11
204	**933**	**0.00**	**0.00**	**351.00**	**825.25**
$\bar{X}$=**12.75**	$\bar{Y}$=**58.31**				

Table 5: Data of Computation for Correlation of TEST and PROD

X (TEST)	Y (PROD)	$X - \bar{X}$	$Y - \bar{Y}$	$(X - \bar{X})2$	$(X - \bar{X})(Y - \bar{Y})$
16	100	3.25	23.06	10.56	74.95
10	46	-2.75	-30.94	7.56	85.08
14	92	1.25	15.06	1.56	18.83
17	100	4.25	23.06	18.06	98.02
14	92	1.25	15.06	1.56	18.83
17	100	4.25	23.06	18.06	98.02
25	69	12.25	-7.94	150.06	-97.23
11	85	-1.75	8.06	3.06	-14.11
6	46	-6.75	-30.94	45.56	208.83
5	31	-7.75	-45.94	60.06	356.02
14	100	1.25	23.06	1.56	28.83
13	100	0.25	23.06	0.06	5.77
13	85	0.25	8.06	0.06	2.02
8	46	-4.75	-30.94	22.56	146.95
11	54	-1.75	-22.94	3.06	40.14
10	85	-2.75	8.06	7.56	-22.17
204	1231	0.00	0.00	351.00	1048.75
$\bar{X}$=12.75	$\bar{Y}$=76.94				

For example of TEST vs. QLTY,
For mean of QLTY ($\bar{Y}$),

$(\bar{Y}) = \frac{69+28+49+72+78+75+60+69+26+43+68+86+69+11+75+55}{16}$ = 58.31

$X - \bar{X}$ = 16- 12.75 = 3.25
$Y - \bar{Y}$ = 69 - 58.31 = 10.69
$(X - \bar{X})^2 = (16- 12.75)^2 = 10.56$
$(X - \bar{X})(Y - \bar{Y}) = 3.25 * 10.69 = 34.73$
Predicted value of Y for external code quality,

$b_1 = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\Sigma(X-\bar{X})^2} = \frac{825.25}{351.00} = 2.35$

$b_0 = \hat{y} - b_1\bar{x}$ = mean ($\bar{Y}$) -2.35 * mean ($\bar{X}$) = 58.31 – (2.35 * 12.75) = 28.34
$\hat{y} = b_0 + b_1x$ = 28.34+ (2.35* 12.75) =58.31
Predicted value of Y for developer productivity,

$b_1 = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\Sigma(X-\bar{X})^2} = \frac{1048.75}{351.00} = 2.99$

$b_0 = \hat{y} - b_1\bar{x}$ = mean ($\bar{Y}$) -2.99 * mean ($\bar{X}$) = 76.94 – (2.99 * 12.75) = 38.8 4
$\hat{y} = b_0 + b_1x$ = 38.84+ (2.99* 12.75) =76.94

5.4. Assessment

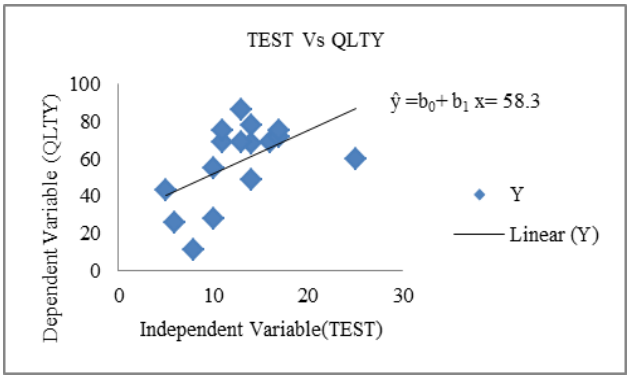


Fig. 4: QLTY on a function of TEST

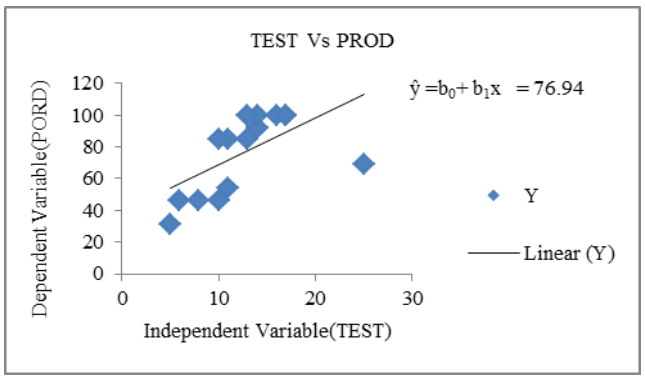


Fig. 5: PROD on a function of TEST

This analytical tool is most often applied to describe data motions over a period of time or correlation between two variables. This system expects that the regression analysis of the information compiled from the

developer productivity by TDD responds positively to questions RQ2. Due to the fact that TDD displays more steps in its process RQ1, a slight decrease in the external code quality is predicted.

In figure 4, the external code quality is slightly decreased by measuring the ordinary least squares (OLS) method of quality (QLTY) [12-13]. In figure 5, the developer productivity is improved by measuring the ordinary least squares (OLS) method of productivity (PROD).

6. Discussion and Comparison of Results

In this section, this paper presents the results acceptance test suite metric of regression analysis. Further, a significant relation between TEST and QLTY, as expressed in RQ1, with a positive was found. Hence, scatter plot figure 2 is an arithmetically expressive relationship between the number of tests and external code quality. Additional, a significant relation between TEST and PORD, as expressed in RQ2, with a somewhat down was found. So, scatterplot figure 3 is an arithmetically expressive correlation between the number of tests and programmer productivity. In this study, the number of tests is a good predictor for TDD programmer productivity [13]. Consequently, developer productivity over the number of tests becomes lightly diminishment and external code quality over the number of tests becomes improvement. Next, this paper presents the results of linear regression analysis. The predicted value of Y ($\hat{y}$) correlation between TEST and QLTY is 58.31. Further, a significant relation between TEST and QLTY, as expressed in RQ1, with a positive linear trend was not found. The linear regression between the two variables is expressed through the equation: QLTY= 28.34 + 2.35* TEST. This equation is plotted in Figure 2. The significance test for the linear regression coefficient, the regression line slope (b_1) is 2.35 and the regression intercept line of y-axis (b_0) is 28.34. Hence there is no arithmetically expressive relationship between the number of tests and external code quality. The predicted value of Y ($\hat{y}$) correlation between the TEST and PROD variables is 76.94. Further, a significant relation between TEST and PORD, as expressed in RQ2, with a positive linear trend was found. The linear regression between the two variables is expressed through the equation: PROD= 38.84 + 2.99* TEST. This equation is plotted in Figure 3. The significance test for the linear regression coefficient, the regression line slope (b_1) is 2.99 and the regression intercept line of y-axis (b_0) is 38.84. Hence there is an arithmetically expressive correlation between the number of tests and programmer productivity [13-14]. In this study, the number of tests is a good predictor for TDD programmer productivity. Consequently, developer productivity becomes improvement and external code quality becomes lightly diminishment.

7. Conclusion

The proposed system is a developing software technology that can support developers to design a code and in their task with resolution. Therefore, the developer will be capable to create extra reliable software. This system has given the developers a more logical accepting of their code and has supported them to advance their development skills. The system counts the bugs and defects over the time-frame. This approach allows thorough unit testing which enhances the quality of the software and advances customer satisfaction. They help with retaining and varying the code. Moreover, the number of acceptance test cases passed and number defects found through static code analysis are used to measure the external code quality. All these measures are consistent with the studies and will be considered as standard measures. When this proposed system assesses the acceptance test suite metric of regression analysis, the result of developer productivity over the number of tests is fewer decreased and the result of external code quality over the number of tests is increased in giving a fixed time-frame. The metric for external code quality and developer productivity used in this system clearly effect. If higher external code quality, lower developer productivity and vice versa, a developer may perhaps be more beneficial by executing as many user stories as possible but dismissing external code quality.

8. Acknowledgments

This research paper is partially supported by academic studies. Professionals were fit to implement more effective with test-driven development. Furthermore, this proposed system observes that the measurement reveal different aspects of a development approach in academic studies.

9. References

[1] Causineou and Chartier, 2010; Outliers Detection and Treatment: a Review, International Journal of Psychological Research, 3(1): 58-67.}

[2] H. Kou, P. M. Johnson, and H. Erdogmus, "Operational definition and automated inference of test-driven development with Zorro," Automated Software Engineering, 2010.

[3] Shaweta Kumar, Sanjeev bansal, "Comparative Study of Test driven Development with Traditional Techniques"; International Journal of Soft computing and Engineering (IJSCE); ISSN:2231-2307,Volume-3, Issue-1, (March 2013).

[4] A.N. Seshu Kumar and S. Vasavi ; "Effective Unit Testing Framework for Automation of Windows Applications"; Aswatha Kumar M.et al.(Eds); Proceedings of ICADC, AISC 174, pp. 813-822. Springerlink .com @ Springer India 2013

[5] Y. Rafique and V. B. Mi˘si´c, "The effects of test-driven development on external quality and productivity: A meta-analysis," IEEE Transactions on Software Engineering, vol. 39, no. 6, pp. 835–856, 2013.

[6] Causevic, A., Shukla, R., & Punnekkat, S. (2013). "Industrial study on test driven development: Challenges and experience" 2013 1st International Workshop on Conducting Empirical Studies in Industry (CESI).

[7] Davide Fucci, Burak Turhan, "On the role of tests in test- driven development: A differentiated and partial replication", Empirical Software Engineering Journal (April 2014, Volume 19, Issue 2, pp 277-302)

[8] Tosun A., Dieste O., Fucci D., Vegas S., Turhan B., Erdogmus H., Santos A., Oivo M., Toro K., Jarvinen J., & Juristo N. An Industry Experiment on the Effects of Test-Driven Development on External Quality and Productivity

[9] Fucci, D., Turhan, B., & Oivo, M. The Impact of Process Conformance on the Effects of Test-driven Development (ESEM2014) 8th Empirical Software Engineering and Measurement, 2014 ACM/IEEE International Symposium on. Turin, Italy.

[10] Fucci, D., Turhan, B., & Oivo, M. On the Effects of Programming and Testing Skills on External Quality and Productivity in a Test-driven Development Context (EASE2015) 19th Evaluation and Assessment in Software Engineering 2015 ACM/IEEE International Conference on., Nanjing, China.

[11] Viktor Farcic , Alex Garcia ; "Java Test-Driven Development"; First published: August 2015; Production reference: 1240815; Published by Packt Publishing Ltd.; Livery Place; 35 Livery Street; Birmingham B3 2PB, UK. ISBN 978-1-78398-742-9; www.packtpub.com; www.it-ebooks.in

[12] Christine Sarikas (GENERAL_EDUCATION) https:// blog. prepscholar.com/independent-and-dependent-variables; Feb 12, 2018.

[13] Tosun, A., Ahmed, M., Turhan, B., & Juristo, N. (2018). On the effectiveness of unit tests in test-driven development. Proceedings of the 2018 International Conference on Software and System Process - ICSSP '18.

[14] Fucci D., Scanniello G., Romano S., Shepperd M., Sigweni B., Uyaguari F., Turhan B., Juristo N., & Oivo M. "An External Replication on the Effects of Test-driven Development Using Blind Analysis" (ESEM2016) 10th Empirical Software Engineering and Measurement 2016 ACM/IEEE International Symposium on., Ciudad Real, Spain.

Chapter 2
Artificial Intelligence and Machine Learning

ISBN 978-981-14-4787-7
Proceedings of 2020 the 10th International Workshop on Computer Science and Engineering (WCSE 2020)
Yangon (Rangoon), Myanmar (Burma), February 26- February 28, 2020, pp. 36-45

Statistical Machine Translation between Myanmar and Myeik

Thazin Myint Oo [1+], Ye Kyaw Thu [2], Khin Mar Soe [1] and Thepchai Supnithi [2]

[1] University of Computer Studies Yangon, Myanmar
[2] National Electronics and Computer Technology Center, Thailand

Abstract. This paper contributes the first evaluation of the quality of machine translation between Myanmar and Myeik (also known as Beik) . We also developed a Myanmar-Myeik parallel corpus (around 10K sentences) based on the Myanmar language of ASEAN MT corpus. In addition, two types of segmentation were studied word and syllable segmentation. The 10 folds cross-validation experiments were carried out using three different statistical machine translation approaches: phrase-based, hierarchical phrase-based, and the operation sequence model (OSM). The results show that all three statistical machine translation approaches give higher and comparable BLEU and RIBES scores for both Myanmar to Myeik and Myeik to Myanmar machine translations. OSM approach achieved the highest BLEU and RIBES scores among three approaches. We also found that syllable segmentation is appropriate for translation quality comparing with word level segmentation results.

Keywords: Statistical machine translation, Myanmar language (Burmese), Myeik dialect, Machine translation for dialects, Parallel corpus developing

1. Introduction

Our main motivation for this research is to investigate SMT performance for Myanmar (Burmese) and Myeik language pair. The Myeik language is closely related to Myanmar (Burmese) language and it is often considered as dialect of Myanmar language. The state-of-the-art techniques of statistical machine translation (SMT) [1], [2] demonstrate good performance on translation of languages with relatively similar word orders [3].

To date, there have been some studies on the SMT of Myanmar language. Ye Kyaw Thu et al. (2016) [4] presented the first large-scale study of the translation of the Myanmar language. A total of 40 language pairs were used in the study that included languages both similar and fundamentally different from Myanmar. The results show that the hierarchical phrase-based SMT (HPBSMT) [5] approach gave the highest translation quality in terms of both the BLEU [6] and RIBES scores [7]. Win Pa Pa et al. (2016) [8] presented the first comparative study of five major machine translation approaches applied to low-resource languages. PBSMT, HPBSMT, tree-to-string (T2S), string-to-tree (S2T) and OSM translation methods to the translation of limited quantities of travel domain data between English and (Thai, Laos, Myanmar) in both directions. The experimental results indicate that in terms of adequacy (as measured by BLEU score), the PBSMT approach produced the highest quality translations. Here, the annotated tree is used only for English language for S2T and T2S experiments. This is because there is no publicly available tree parser for Lao, Myanmar and Thai languages. According to our knowledge, there is no publicly available tree parser for Myeik language and thus we cannot apply S2T and T2S approaches for Myanmar-Myeik language pair. From their RIBES scores, we noticed that OSM approach achieved best machine translation performance for Myanmar to English translation. Moreover, we learned that OSM approach gave highest translation performance translation between Khmer (the official language of Cambodia) and twenty other languages, in both directions [9].

[+] Corresponding author. . Tel.: +9595072165
E-mail address: thazinmyintoo@ucsy.edu.mm

Based on the experimental results of previous works, in this paper, the machine translation experiments were carried out using PBSMT, HPBSMT and OSM.

2. Related Work

Karima Meftouh et al. built PADIC (Parallel Arabic Dialect Corpus) corpus from scratch, then conducted experiments on cross dialect Arabic machine translation [10]. PADIC is composed of dialects from both the Maghreb and the Middle-East. Some interesting results were achieved even with the limited corpora of 6,400 parallel sentences.

Using SMT for dialectal varieties usually suffers from data sparsity, but combining word-level and character-level models can yield good results even with small training data by exploiting the relative proximity between the two varieties [11]. Friedrich Neubarth et al. described a specific problem and its solution, arising with the translation between standard Austrian German and Viennese dialect. They used hybrid approach of rule-based preprocessing and PBSMT for getting better performance.

Pierre-Edouard Honnet et al. proposed solutions for the machine translation of a family of dialects, Swiss German, for which parallel corpora are scarce [12]. They presented three strategies for normalizing Swiss German input in order to address the regional and spelling diversity. The results show that character-based neural MT was the most promising one for text normalization and that in combination with PBSMT achieved 36% BLEU score.

3. Myeik Language

The Myeik dialect is a dialect of Burmese that is spoken in Myeik (Beik), a town situated in the southern part of Tanintharyi Division (around 12Â°25'N, 98Â°37'E), Republic of the Union of Myanmar[13]. Myeik dialect is one of the southernmost dialects of Burmese and can be regarded as the southernmost distribution of the Tibeto-Burman languages. Myeik was formerly called Mergui in English.

Myeik dialect has peculiar characteristics in terms of tonal contours, and voice quality in the tones and vowels. The tone of this dialect, which corresponds to the Standard Burmese creaky falling tone, has a rising contour and is pharyngealized [14]. Vowels of the syllables corresponding to Standard Burmese stopped syllables are pronounced with a conspicuous creaky phonation. Previous studies have paid little attention to these facts. Tones and his peculiar to this dialect are also described in this paper [15]. Dialogues cover as many as possible of the most basic grammatical items of Burmese, translating them into the Myeik dialect can be the basis for future studies of morphosyntactic phenomena of this dialect [16] .

There are some examples of myeik and Myanmar.

bk : မင်း ငါ့ ကို ကြေးပြား ပေး ဝို့ မေ့ နေရယ်လား။
my: မင်း ငါ့ ကို ပိုက်ဆံ ပေး ဖို့ မေ့ နေပြီလား ။
(“Do you forget paying money to me.” in English)

bk : ငါ မောလင်း နိုင်ငံခြား သော မယ်။
my : ကျွန်တော် မနက်ဖြန် နိုင်ငံခြား သွား မယ် ။
(“I will go foreign tomorrow .” in English)

bk : ကျွန်တော် ဒယ် ဝို လာ ရဇာ ပျော် ရယ် ။
my : ကျွန်တော် ဒီ လာ ရတာ ပျော် တယ် ။
(“ I am happy to come here.” in English)

In the above examples, the underlined words that have same meaning but have different spellings such as “ကြေးပြား ” vs “ပိုက်ဆံ” (“money” in English), “မောလင်း” vs “မနက်ဖြန်” (“tomorrow” in English), “ဒယ်” vs “ဒီ ” (“this” in English).

4. Methodology

In this section, we describe the methodology used in the machine translation experiments for this paper.

4.1. Phrase-Based Statistical Machine Translation

A PBSMT translation model is based on phrasal units [1]. Here, a phrase is simply a contiguous sequence of words and generally, not a linguistically motivated phrase. A phrase-based translation model typically gives better translation performance than word-based models. We can describe a simple phrase-based translation model consisting of phrase-pair probabilities extracted from corpus and a basic reordering model, and an algorithm to extract the phrases to build a phrase-table [17].

The phrase translation model is based on noisy channel model. To find best translation e that maximizes the translation probability P(e|f) given the source sentences; mathematically. Here, the source language is French and the target language is an English. The translation of a French sentence f into an English sentence e is modeled as equation 1.

$$e = \text{argmax}_e P(e|f) \quad (1)$$

The final mathematical formulation of phrase-based model is as follows:

$$\text{argmax e } P(e|f) = \text{argmax e } P(f|e)\ P(e) \quad (2)$$

We note that denominator P(e|f) can be dropped because for all translations the probability of the source sentence remains the same . The P(e|f) variable can be viewed as the bilingual dictionary with probabilities attached to each entry to the dictionary (phrase table). The P(e) variable governs the grammaticality of the translation and we model it using n-gram language model under the PBMT paradigm.

4.2. Hierarchical Phrase-Based Statistical Machine Translation

The hierarchical phrase-based SMT approach is a model based on synchronous context-free grammar [5]. The model is able to be learned from a corpus of unannotated parallel text. The advantage this technique offers over the phrase-based approach is that the hierarchical structure is able to represent the word re-ordering process. The re-ordering is represented explicitly rather than encoded into a lexicalized re-ordering model (commonly used in purely phrase-based approaches). This makes the approach particularly applicable to language pairs that require long-distance re-ordering during the translation process [18]. An example of hierarchical phrase-based grammar rules between Myanmar and Myeik from a HPBSMT model is as follows:

ငါ တွေးစာ တအား [X] ||| ကျွန်တော့် အတွေးနဲ့ [X]
ငါ တွေးစာ တအား [X][X] [X] ||| ကျွန်တော့် အတွေးနဲ့
ငါ တွေးစာ တအား တူ [X] ||| ကျွန်တော့် အတွေးနဲ့ တူ [X]
ငါ တွေးစာ တအား တူ ရယ် [X] ||| ကျွန်တော့် အတွေးနဲ့ တူ တယ် [X]

4.3. Operation Sequence Model

The operation sequence model that can combines the benefits of two state-of-the-art SMT frameworks named n-gram-based SMT and phrase-based SMT. This model simultaneously generate source and target units and does not have spurious ambiguity that is based on minimal translation units [19][20]. It is a bilingual language model that also integrates reordering information. OSM motivates better reordering mechanism that uniformly handles local and non-local reordering and strong coupling of lexical generation and reordering. It means that OSM can handle both short and long distance reordering. The operation types are such as generate, insert gap, jump back and jump forward which perform the actual reordering. The following shows an example translation process of English sentence “Please sit here” into Myanmar language with the OSM.

Source: Please sit here
Target: ကျေးဇူးပြုပြီး ဒီမှာ ထိုင်
Operation 1: Generate (Please, ကျေးဇူးပြုပြီး)
Operation 2: Insert Gap

Operation 3: Generate (here, ဒီမှာ)
Operation 4: Jump Back (1)
Operation 5: Generate (sit, ထိုင်)

5. Experiments

5.1. Statistics

We used 10K Myanmar sentences (without name entity tags) of the ASEAN-MT Parallel Corpus [21], which is a parallel corpus in the travel domain. It contains six main categories and they are people (greeting, introduction and communication), survival (transportation, accommodation and finance), food (food, beverage and restaurant), fun (recreation, traveling, shopping and nightlife), resource (number, time and accuracy), special needs (emergency and health). Manual translation into Myeik language was done by native Myeik students from Computer University (Myeik). Word segmentation for Myeik was done manually and there are exactly 68,035 words in total. We held 10-fold cross-validation experiments and used 7,867 to 7,893 sentences for training, 1,389 to 1,393 sentences for development and 1,014 to 1,044 sentences for evaluation respectively.

5.2. Word Segmentation

In both Myanmar and Myeik text, spaces are used for separating phrases for easier reading. It is not strictly necessary, and these spaces are rarely used in short sentences. There are no clear rules for using spaces, and thus spaces may (or may not) be inserted between words, phrases, and even between a root words and their affixes. Although Myanmar sentences of ASEAN-MT corpus is already segmented, we have to consider some rules for manual word segmentation of Myeik sentences. We defined Myeik "word" to be meaningful units and affix, root word and suffixe(s) are separated such as "စား ရယ်". Here, "စား" ("eat" in English) is a root word and suffix "ရယ်" . Similar to Myanmar language, Myeik plural nouns are identified by following particle. We also put a space between noun and the following particle, for example a Myeik word "သားကင်းငယ်တွေ" (children) is segmented as two words "သားကင်းငယ်" and the particle "တွေ". In our manual word segmentation rules, compound nouns are considered as one word and thus, a compound word "ကြေးပြား + အိတ်" ("money" + "bag" in English) is written as one word "ကြေးပြားအိတ်" ("wallet" in English). Myeik adverb words such as "အား" ("very" in English) also considered as one word. The following is an example of word segmentation for a sentence in our corpus and the meaning is "why are you beaten the children."

Unsegmented sentence:
ဘာဖြစ်ရိသားကင်းငယ်တွေကိုရိုက်နေရယ်။

Segmented sentence:
ဘာဖြစ်ရိ သားကင်းငယ် တွေ ကို ရိုက် နေရယ်။

In this example, **"သားကင်းငယ်တွေ "** ("children" in English) is a compound word of **"သားကင်းငယ်"** ("child" in English) and a particle **"တွေ"** are segmented as two words. A root word **"ရိုက်"** and the suffix **"နေရယ်"** are also segmented as two words **" ရိုက် နေရယ်"** ("are beating" in English).

5.3. Syllable Segmentation

Generally, Myanmar words are composed of multiple syllables, and most of the syllables are composed of more than one character. Syllables are composed of Myanmar words. If we only focus on consonant-based syllables, the structure of the syllable can be described with Backus normal form (BNF) as follows:

Syllable := CMV[CK][D]

Here, "C" stands for consonants, "M" for medials, "V" for vowel, "K" for vowel killer character, and "D" for diacritic characters. Myanmar syllable segmentation can be done with a rule-based approach, finite state automation (FSA) or regular expressions (RE) (https://github.com/ye-kyawthu/sylbreak). In our

experiments, we used RE based Myanmar syllable segmentation tool named. The following is an example of syllable segmentation for a Myeik sentence in our corpus and the meaning is

Unsegmented Myeik sentence:

bk: ဘာဖြစ်ရိသားကင်းယေ်တွေကိုရိုက်နေရယ်။

Syllable Segmented Myeik sentence:

bk: ဘာ ဖြစ် ရိ သား ကင်း ယေ် တွေ ကို ရိုက် နေ ရယ် ။

5.4. Moses SMT System

We used the PBSMT, HPBSMT and OSM system provided by the Moses toolkit [2] for training the PBSMT, HPBSMT and OSM statistical machine translation systems. The word segmented source language was aligned with the word segmented target language using GIZA++ [22]. The alignment was symmetrize by grow-diag-final and heuristic [1]. The lexicalized reordering model was trained with the msd-bidirectional-fe option [23]. We use KenLM [24] for training the 5-gram language model with modified Kneser-Ney discounting [25]. Minimum error rate training (MERT) [26] was used to tune the decoder parameters and the decoding was done using the Moses decoder (version 2.1.1) [2]. We used default settings of Moses for all experiments.

6. Evaluation

We used two automatic criteria for the evaluation of the machine translation output. One was the de facto standard automatic evaluation metric Bilingual Evaluation Understudy (BLEU) [6] and the other was the Rank-based Intuitive Bilingual Evaluation Measure (RIBES) [7]. The BLEU score measures the precision of n-gram (over all $n \leq 4$ in our case) with respect to a reference translation with a penalty for short translations [6]. Intuitively, the BLEU score measures the adequacy of the translation and large BLEU scores are better. RIBES is an automatic evaluation metric based on rank correlation coefficients modified with precision and special care is paid to word order of the translation results. The RIBES score is suitable for distance language pairs such as Myanmar and English. Large RIBES scores are better.

7. Results and Discussion

The BLEU and RIBES score results for machine translation experiments with PBSMT, HPBSMT and OSM are shown in Table 1. Bold numbers indicate the highest scores among three SMT approaches. The RIBES scores are inside the round brackets. Here, “my” stands for Myanmar, “bk” stands for Myeik, “src” stands for source language and “tgt” stands for target language respectively.

Table 1: Average BLEU and RIBES scores for PBSMT, HPBSMT and OSM using word segmentation (Evaluation with syllable unit)

src-tgt	PBSMT	HPBSMT	OSM
bk-my	44.12 (0.87488)	44.07 (0.87513)	**44.33 (0.87531)**
my-bk	33.25 (0.84045)	33.33 (0.83882)	**33.41 (0.83991)**

The BLEU and RIBES score results for machine translation experiments with PBSMT, HPBSMT and OSM between Myanmar and Myeik languages are shown in Table 1. To compare with syllable results, the translation results were decomposed into their constituent syllables to ensure that the results are cross-comparable. From the results, OSM method achieved the highest BLEU and RIBES score for both Myanmar-Myeik and Myeik-Myanmar machine translations. Interestingly, the BLEU and RIBES score of all three methods are comparable performance. Our results with current parallel corpus indicate that Myeik to

Myanmar machine translation is better performance (around 10 BLEU and 0.04 RIBES scores higher) than Myanmar to Myeik translation direction. Our results with syllable segmentation also indicate that Myeik to Myanmar machine translation is better performance (around 15 BLEU and 0.03 RIBES score higher) than Myanmar to Myeik translation direction. Our investigation clearly show that getting the higher scores with syllable segmentation for bi-directional Myanmar to Myeik machine translation.

As we expected, generally, machine translation performance of all three SMT approaches between Myanmar and Myeik languages achieved comparable scores for both BLEU and RIBES. The reason is that as we mentioned in Section 3, the two languages, Myanmar and Myeik are very close languages.

Table 2: Average BLEU and RIBES scores for PBSMT, HPBSMT and OSM using syllable segmentation

src-tgt	PBSMT	HPBSMT	OSM
bk-my	70.017 (0.95728)	69.894 (0.95656)	**70.545** **(0.95793)**
my-bk	54.606 (0.92213)	54.404 (0.92194)	**55.112** **(0.92315)**

8. Error Analysis

The top 10 confusion pairs of OSM model for for Myeik-Myanmar machine translation with word segmentation are shown in table 3.

Table 3: Top 10 confusion pairs of OSM model for Myeik-Myanmar machine translation with word segmentation

Freq	Confusion Pair (REF→HYP)
45	ဝို ==> ကို
35	မင်း ==> နင်
23	ကို ==> ဝို
15	သူ ==> ဒယ်ကောင်မငယ်
14	မင့် ==> နင်
12	ငါ ==> ကျွန်တော်
12	နင် ==> ခင်ဗျား
12	လဲ ==> ရို
11	သွား ==> သော
8	ဝ ==> ရ

We also made manual error analysis on translated outputs of the best OSM model, and we found that dominant errors are different in sentence level. We will introduce four frequent error patterns and they are “Male-Female Vocabulary Error”, “Paraphrasing Error”, “Word Segmentation Error” and “Negative Error”. The followings are some example translation mistakes of Myanmar-Myeik machine translation for each category:

Male-Female Vocabulary Error

SOURCE: သူမ က သူ့ ကို အပြစ်တင် တယ် ။
Scores: (#C #S #D #I) 3 3 0 1
REF: ****** ဒယ်ကောင်မငယ် ဟ သူ့ဝို အပြစ်တင် ရယ် ။
HYP: သူ က သူ့ ကို အပြစ်တင် ရယ် ။
Eval: I S S S

SOURCE: အဲ့ဒါ ကို သူမ မှတ်မထား ဘူးလား ။
Scores: (#C #S #D #I) 3 2 0 1
REF: ********** ဒယ်စာဝို ဒယ်ကောင်မငယ် မှတ်မထား ရလား ။
HYP: ဒယ်ဇာ ကို သူ မှတ်မထား ရလား ။
Eval: I
S S

SOURCE: သူမ အရမ်း စိတ်အားထက်သန် နေတယ် ။
Scores: (#C #S #D #I) 2 3 0 0
REF: သူလေ တအား စိတ်အားထက်သန် နေရယ် ။
HYP: ဒယ်ကောင်မငယ် အလွန် စိတ်အားထက်သန် ရယ် ။
Eval: S S S

Paraphrasing Error

SOURCE: အကြင်နာ ရော ရှိ ရဲ့လား ။
Scores: (#C #S #D #I) 3 2 0 0
REF: အကြင်နာ ကော ရှိ ရယ်ပဲ့လား ။
HYP: အကြင်နာ ရော ရှိ ပဲ့လား ။
Eval: S S

SOURCE: သူ့ကိုသူ အားမပေး ချင်ဘူး ဟုတ်လား ။
Scores: (#C #S #D #I) 2 3 1 1
REF: ********** သူ့ဝိုသူ အားမပေး ရမော် ဟုတ် ဝယ်မှန်း ။
HYP: သူ့ကို သူ အားမပေး ********** ချင်ရမော် ဟုတ်ဝယ်လား ။
Eval: I S D S S

SOURCE: အတ္ထုပတ္တိ တွေ ဘယ်နားမှာ တွေ့ နိုင်လဲ ကျေးဇူးပြုပြီး ပြောပြ ပါလား ။
Scores: (#C #S #D #I) 3 5 0 1
REF: အတ္ထုပတ္တိ ဒေ ********** ဘယ်နားမှာ တွေ့နိုင်လဲ ကျေးဇူးပြုပြီး ပြောပြ နိုင်လား ။
HYP: အတ္ထုပ္ပတ္တိ ဒေ ဘယ်မှာ တွေ့နိုင် ရယ် ကျေးဇူးပြုပြီး ပြော ပြ ။
Eval: S I S S S
S

Word Segmentation Error

SOURCE: ခင်ဗျား အဲ့ဒါ ကို ချီးကျူးချင်ချီးကျူး မချီးကျူး ချင်နေ ။
Scores: (#C #S #D #I) 3 3 0 0
REF: မင့် အဲဇာဝို ချီးကျူးချင်ချီးကျူး မချီးကျူး ချင်နေ ။
HYP: မင့် အဲဇာဝို ချီးကျူး ချင်ချီးမွမ်း မချီးမွမ်းချင်နေ ။
Eval: S S

SOURCE: သူမ ကို တသက်လုံး ခွဲ သွား မှာ မ ဟုတ် ဘူး ။
Scores: (#C #S #D #I) 4 3 3 0
REF: ဒယ်ကောင်မငယ် ဝို တသက်လုံး ခွဲ သော မှာ မ ဟုတ် ဝ ။
HYP: ဒယ်ကောင်မငယ် ကို တသက်လုံး ခွဲ ********* ********* *** သွားမှာ ဟုတ်ဝ ။
Eval: S D D D S S

Negative Error

SOURCE: သူမ ငို မှာ မ ဟုတ် ဘူး ။
Scores: (#C #S #D #I) 2 2 3 0
REF: သူ ငို မှာ မ ဟုတ် ဝ ။
HYP: ဒယ်ကောင်မငယ် ငို ********* *** ********** ဟုတ်ဝ ။
Eval: S D D D S

SOURCE: သူမ စကား မ ပြော ဘူး ။
Scores: (#C #S #D #I) 2 2 2 1
REF: ********** ဘယ်ဒယ်ကောင်မငယ် စကား မ ပြော ဘူး ။
HYP: အပြင်မှာ ဘယ်ဒယ်ကောင်မငယ် ********** *** စကားပြော ဟုတ်ဝ ။
Eval: I D D S S

SOURCE: ခင်ဗျား အတင်းဝင် ရမယ် မ ဟုတ် လား ။
Scores: (#C #S #D #I) 4 1 0 2
REF: ခင်ဗျား အတင်းဝင် ရမယ် *** ********** ဟုတ်ဝ ။
HYP: ခင်ဗျား အတင်းဝင် ရမယ် မ ဟုတ် ဝ ။
Eval: I I S

"SOURCE" is the test sentence of Myanmar language, "Scores" are operation scores of the Edit Distance , "C" is the number of correct words, "S" is the number of substitutions, "D" is the number of deletions, "I" is the number of insertions, "REF" for reference (i.e. Myeik sentence), "HYP" for hypothesis and "Eval" is the ordered sequence of edit operations. We found that translation error of male to female vocabulary and vice versa happen between Myeik-Myanmar translation such as "ဒယ်ကောင်မငယ်" ("she" in English) to "သူ" ("he" in English). The second category, paraphrasing errors are really interesting and it is also proved that two language are similar languages. In our paraphrasing error examples, the meanings of all reference and hypothesis pairs are the same. Some errors are just the difference between the formal (polite form) and informal written form such as "ရှိရယ်ပဲ့လား" (polite form of ending phrase "ရှိပဲ့လား" in Myeik conversation) and "ရှိလား". One of the possible reasons for the word segmentation errors is inconsistent word segmentation of human translators such as "ချီးကျူးချင်ချီးကျူး" and "ချီးကျူး ချင်ချီးကျူး" ("admirably" in English). We also found that one more frequent translation errors between Myeik-Myanmar and Myanmar-Myeik machine translation is changing into negative form (e.g. "စကားပြော" ("to speak" in English) and "စကားမပြော" ("no speaking" in English).

9. Conclusion

This paper contributes the first PBSMT, HPBSMT and OSM machine translation evaluations from Myanmar to Myeik and Myeik to Myanmar. We used the 10K Myanmar-Myeik parallel corpus that we constructed to analyze the behavior of a dialectal Myanmar-Myeik machine translation with word and syllable segmentation unit. The result get better translation result in syllable translation unit than word level. We showed that higher BLEU and RIBES scores can be achieved for Myeik-Myanmar language pair even with the limited data. This paper also present detail analysis on confusion pairs of machine translation

between Myanmar-Myeik and Myeik-Myanmar. In the future we plan to test PBSMT, HPBSMT and OSM models with other Myanmar dialect languages such as Yaw and Danu.

10.Acknowledgements

We would like to express our gratitude to all students of Myanmar-Myeik translation team namely, Aung Win Htut, Aung Thurin Tun, Nandar Win, Myat Hein Tun, Aye Thet Moe, Yadanar Moe, Paing Paing Tun, Shwe Yi Oo, Ei Ei Hiwe, Hnin Sett Pwint Paing, Zaw Zaw Aung, Zin Pwint Htwe and Hnin Wutyi Oo for translation between Myanmar and myeik sentences. Last but not least, we would like to thank Daw Thandar Win (Prorector, University of Computer Studies Myeik) for all the help and support during our stay at University of Computer Studies Myeik.

11.References

[1] P. Koehn, F. J. Och, and D. Marcu, “Statistical phrase-based translation.” in Proc. of HTL-NAACL, 2003, pp. 48–54.

[2] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, “Moses: Open source toolkit for statistical machine translation.” in Proc. of ACL, 2007, pp. 177–180.

[3] P. Koehn, “Europarl: A parallel corpus for statistical machine translation.” in Proc. of MT summit, 2005, pp. 79–86.

[4] Ye Kyaw Thu, Andrew Finch, Win Pa Pa, and Eiichiro Sumita, “A Large-scale Study of Statistical Machine Translation Methods for Myanmar Language”, in Proc. of SNLP2016, February 10-12, 2016.

[5] Chiang, D., “Hierarchical phrase-based translation”, Computational Linguistics 33(2), 2007, pp. 201-228.

[6] Papineni, K., Roukos, S., Ward, T., Zhu, W., “BLEU: a Method for Automatic Evaluation of Machine Translation”, IBM Research Report rc22176 (w0109022), Thomas J. Watson Research Center, 2001

[7] Isozaki, H., Hirao, T., Duh, K., Sudoh, K., Tsukada, H, “Automatic evaluation of translation quality for distant language pairs”, in Proc. of the 2010 Conference on Empirical Methods in Natural Language Processing, pp. 944-952.

[8] Win Pa Pa, Ye Kyaw Thu, Andrew Finch and Eiichiro Sumita, "A Study of Statistical Machine Translation Methods for Under Resourced Languages", 5th Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU Workshop), 09-12 May, 2016, Yogyakarta, Indonesia, Procedia Computer Science, Volume 81, 2016, pp. 250–257.

[9] Ye Kyaw Thu, Vichet Chea, Andrew Finch, Masao Utiyama and Eiichiro Sumita, "A Large-scale Study of Statistical Machine Translation Methods for Khmer Language", 29th Pacific Asia Conference on Language, Information and Computation, October 30 - November 1, 2015, Shanghai, China, pp. 259-269.

[10] Karima Meftouh, Salima Harrat, Salma Jamoussi, Mourad Abbas and Kamel Smaili, “Machine Translation Experiments on PADIC: A Parallel Arabic DIalect Corpus”, oin Proc. of the 29th Pacific Asia Conference on Language, Information and Computation, PACLIC 29, Shanghai, China, October 30 - November 1, 2015, pp. 26-34.

[11] Neubarth Friedrich, Haddow Barry, Huerta Adolfo Hernandez and Trost Harald, “A Hybrid Approach to Statistical Machine Translation Between Standard and Dialectal Varieties”, Human Language Technology, Challenges for Computer Science and Linguistics: 6th Language and Technology Conference, LTC 2013, Poznan, Poland, December 7-9, 2013, Revised Selected Papers, pp. 341–353.

[12] Pierre-Edouard Honnet, Andrei Popescu-Belis, Claudiu Musat and Michael Baeriswyl, “Machine Translation of Low-Resource Spoken Dialects: Strategies for Normalizing Swiss German”, CoRR journal, volume (abs/1710.11035), 2017.

[13] Wikipedia of Myeik:

https://en.wikipedia.org/wiki/Myeik_dialect

https://en.wikipedia.org/wiki/Myeik,_Myanmar

[14] Bradley, David. 1982. Register in Burmese. (In) D. Bradley (ed.) “Papers in South-East Asian Linguistics” No. 8:

Tonation. Pacific Linguistics Series No. 62, pp. 117-132.

[15] John Okell , "Three Burmese Dialects", 1981, London Oxford University press, Univeristy of London.

[16] Khin Pale 1974. "A study of Myeik daily vocabulary" , B.A.term paper, Mawlamyaing University, Myanmar

[17] Lucia Specia, "Tutorial, Fundamental and New Approaches to Statistical Machine Translation", International Conference Recent Advances in Natural Language Processing, 2011

[18] Braune, Fabienne and Gojun, Anita and Fraser, Alexander, "Long-distance reordering during search for hierarchical phrase-based SMT", in Proc. of the 16th Annual Conference of the European Association for Machine Translation, 2012, Trento, Italy, pp. 177-184.

[19] Durrani, Nadir and Schmid, Helmut and Fraser, Alexander, "A Joint Sequence Translation Model with Integrated Reordering", in Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, 2011, Portland, Oregon, pp. 1045-1054.

[20] Nadir Durrani, Helmut Schmid, Alexander M. Fraser, Philipp Koehn and Hinrich Schutze, "The Operation Sequence Model - Combining N-Gram-Based and Phrase-Based Statistical Machine Translation", Computational Linguistics, Volume 41, No. 2, 2015, pp. 185-214.

[21] Prachya, Boonkwan and Thepchai, Supnithi, "Technical Report for The Network-based ASEAN Language Translation Public Service Project", Online Materials of Network-based ASEAN Languages Translation Public Service for Members, NECTEC, 2013

[22] Och Franz Josef and Ney Hermann, "Improved Statistical Alignment Models", in Proc. of the 38th Annual Meeting on Association for Computational Linguistics, Hong Kong, China, 2000, pp. 440-447.

[23] Tillmann Christoph, "A Unigram Orientation Model for Statistical Machine Translation", in Proc. of HLT-NAACL 2004: Short Papers, Stroudsburg, PA, USA, 2004, pp. 101-104.

[24] Heafield, Kenneth, "KenLM: Faster and Smaller Language Model Queries", in Proc. of the Sixth Workshop on Statistical Machine Translation, WMT '11, Edinburgh, Scotland, 2011, pp. 187-197.

[25] Chen Stanley F and Goodman Joshua, "An empirical study of smoothing techniques for language modeling", in Proc. of the 34th annual meeting on Association for Computational Linguistics, 1996, pp. 310-318.

[26] Och Franz J., "Minimum error rate training in statistical machine translation", in Proc. of the 41st Annual Meeting n Association for Computational Linguistics – Volume 1,Association for Computer Linguistics, Sapporo, Japan, July, 2003, pp.160-167.

ISBN 978-981-14-4787-7
Proceedings of 2020 the 10th International Workshop on Computer Science and Engineering (WCSE 2020)
Yangon (Rangoon), Myanmar (Burma), February 26- February 28, 2020, pp. 46-51

ICD-10 Auto-coding System Using Deep Learning

Ssu-Ming Wang [1], Feipei Lai [1,2+], Chang-Sung Sung [2] and Yang Chen [2]

[1] Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan Univ., Taipei, Taiwan
[2] Department of Computer Science and Information Engineering, National Taiwan Univ., Taipei, Taiwan

Abstract. In this research, we aim to construct an automatic ICD-10 coding system. ICD-10 is a medical classification standard which is strongly related to scope of payment in health insurance. However, the work of ICD-10 coding is time-consuming and tedious to ICD coders. Therefore, we build an ICD-10 coding system based on NLP approach to reduce their workload. The result of f1-score in whole label prediction task is up to 0.67 and 0.58 in CM and PCS, respectively. In addition, recall@20 in whole label prediction task is up to 0.87 and 0.81 in CM and PCS, respectively. In the future, we will keep working on combining the current work with the rule-based coding system and applying the other brand new NLP techniques to improve our performance.

Keywords: Deep learning, Deep Neural Network, Natural Language Processing (NLP), ICD-10

1. Introduction

Auto-coding for ICD-10 (The International Statistical Classification of Diseases and Related Health Problems 10th Revision, ICD-10) based on free-text electronic health records (EHR) has drawn great attention in the field of clinical management system. The target of this research is to construct an automatic ICD-10 coding system via Natural Language Processing (NLP) technology.

The ICD-10 is a medical classification list released by World Health Organization (WHO) which defines the universe of diseases, disorders, injuries and other related health conditions and the classifying standard of diagnosis [1]. Since the first publication in 1893, ICD was widely used in fields such as health insurance.

After a statistical processing, the disease classification data can be applied to the clinical management system or be an evaluation factor for the health care quality. Also, since the bureau of national health Insurance, Taiwan started to use ICD code as a reference when evaluating the amount of premium subsidies in the diagnosis-related group prospective payment system, ICD codes have become one of the most important index for the hospital to apply for reimbursement and subsidy [2].

Currently, the reference material for ICD coding are mainly unstructured, i.e. free-text data. Different from structured data, unstructured data would not have fixed format and clear rules such as column length or type. The most common expression type of such data is free text included disease description, history, or diagnosis records of patients which are difficult to define the storage formality strictly. Hence, traditionally, the classification of ICD code was mainly relying on the person who has read a plenty of clinical language documents to handle the complicated procedure of ICD-10 classification. The wide and detail scope of this classification method and the need of referring to the classification rules and medical literature make the classifying task time-consuming and tedious, even the most professional staff in this field takes lots of efforts at finishing such classification. Therefore, an automatic ICD-10 coding system based on Deep Neural Network (DNN) model by applying supervised learning [3] illustrated in Figure 1 is proposed for providing assistance to ICD-10 coders.

+ Corresponding author. Tel.: + (8862) 3366-4888 #419; fax: +(8862) 2362-8167.
E-mail address: flai@ntu.edu.tw.

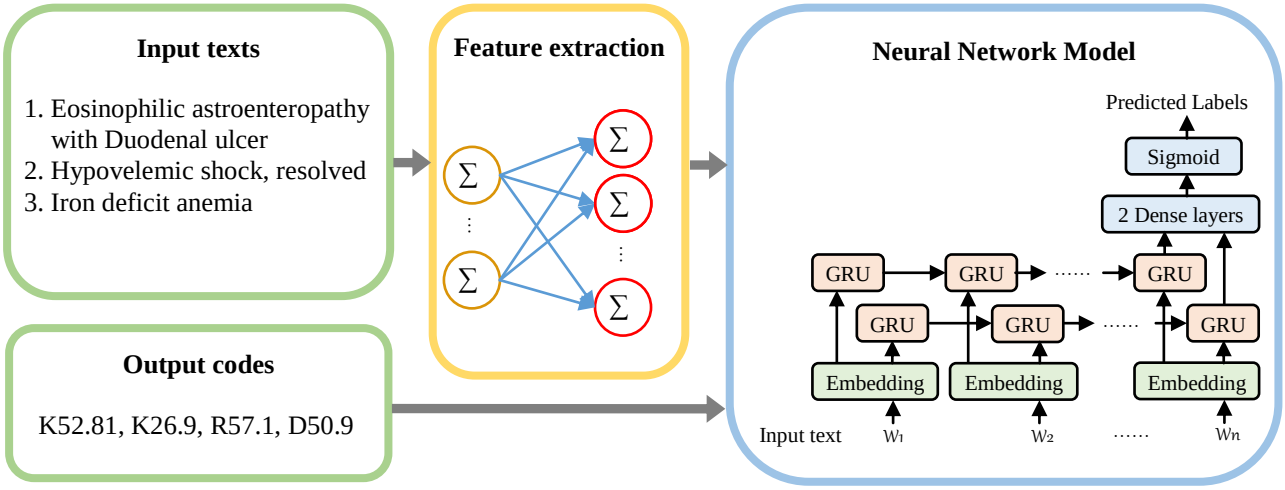


Fig. 1: Training pipeline of ICD-10 classification model.

2. Method

2.1. Data Description

Data including chief complaint, pathology report, physical examination, progress, history, transferring out of ICU diagnosis, and discharge diagnosis between January 2016 to July 2017 was acquired from patients in National Taiwan University Hospital (NTUH). The complete data are about 250,000 records in total and includes up to 14,602 ICD-10 labels which can be divided into 21 classes according to the classes of diseases such as nervous system, immune system, etc. The data will be split to 90% and 10% as training and validation sets.

2.2. Preprocessing

To ensure the uniqueness of the patient account identity and data record, duplicated and null records will be dropped and merged by account identity. Then, procedure including punctuation marks eliminating, case converting, stop words removal, and typos correction are applied for training more effective word embedding model and accelerating the training process. Word tokenization and sequence padding are based on Keras (version 2.2.4) tokenizer.

2.3. Feature Extraction

In this research, Word2Vec, a NLP technique, was applied to transform free-text words to numerical vectors by Continuous Bags of Words (CBOW) model [4]. In CBOW algorithm, Words of sentences were one-hot encoded to binary matrix and then trained with the fully-connected neural network. The previous output eventually pass a softmax classifier, giving the nearby word, i.e. the words within the giving window size, appearing probability for loss computing. Within the process, the hidden layer weight matrix can be extracted to build the word embedding matrix and the corresponding word dictionary.

2.4. Deep Neural Network model

The classification model constructed with four neural network layers, including Recurrent Neural Network (RNN) and Fully-Connected Neural Network (Dense), is shown in Table 1. The first layer is word embedding layer, which transforms the tokenized word list input into word vectors. The second layer is a bidirectional Gated Recurrent Unit (BiGRU) layer [5]. The gating mechanism of GRU solves the vanishing gradient problem that sometimes comes with the standard RNN. In comparison, GRU consumes less time for convergence than the Long Short-Term Memory (LSTM) [6]. The remaining two layers are Dense layers, where the final dense layers should output the vector with the dimension we expect to predict. In our case, there are 21 chapters of ICD-10 and 14602 labels in NTUH data records in total makes the final dense layer size set to 21-dimension and 14602 dimension separately. Each dimension indicates the probability of the code is associated with the diagnosis input.

Table 1: Hyperparameters of whole label classification model.

Hyperparameters	Size
Embedding layer	300
BiGRU layer	256
Dense layer 1	1024
Dense layer 2	14602
Dropout	0.4

2.5. Score Metric

F1-score is the harmonic mean of recall and precision, which are the number of correct positive results divided by the number of all positive results returned by the classifier and the number of correct positive results divided by the number of all relevant samples, hence, appropriate for evaluating the performance of a multi-label classification task. For the realistic application in ICD-10 auto-coding system, recall@20, which calculates the probability of correct answers in first 20 predicting result returned by classifier, is also applied for validating the model performance.

2.6. ICD-10 Predicting Interface

An ICD-10 auto-coding system prototype was built based on python3, ASP.NET Core 2.0 MVC, and Vue.js. The frontend discharge diagnosis input will call for the Web API and sent the case information to the backend. The well-trained DNN model will then be executed with python3 and return the predicting results, providing disease coders the top 20 related ICD-10-CM and ICD-10-PCS codes for auxiliary.

3. Result And Discussion

3.1. ICD-10 Chapter Classification

ICD-10 CM codes are composed of 3 to 7 characters and can be divided into 22 chapters as Table 2 21 classes and U00 to U99. Each of chapter has its title presenting the disease. For example, A00-B99 is about “Certain infectious and parasitic diseases”. Considering that U00-U99 blocks being about “Codes for special purposes” are not related to diseases, our model does not predict the ICD-10 codes between U00 to U99. Hence, in ICD-10 chapter classification task, the ICD-10 codes corresponding each diagnosis record were formatted to 3 characters by removing the last 0 to 4 characters for 21 categories training and predicting. The validation performance is shown in Figure 2.

As Figure 2 shows, discharge diagnosis can achieve the best performance on f1-score of 0.86 on the average of 21 chapters and achieve over 0.5 on f1-score in H60 to H95 and P00-P96 blocks, which only have 1,820 and 2,275 samples in our dataset. Hence, the discharge diagnosis was chosen as the ICD-10 auto – coding reference, i.e. the training data, in whole label classification task and ICD-10 auto-coding system.

3.2. ICD-10 CM Whole Label Classification

According to the ICD-10 chapter classification outcome, we use discharge diagnosis as training data in ICD-10 CM whole label classification task. In NTUH dataset, the complete ICD-10-CM codes, i.e. CM codes with 3 to 7 characters, corresponding to discharge diagnosis records are 14,602 labels. Comparing to the previous study on ICD-9 classification with 85,522 training data and f1-score 0.41 [7], with 300 as embedding dimension, our DNN classification model achieve 0.67 on f1-score and 0.86 on recall@20 as shown in Figure 3.

The model performance in each NTUH division is also tested in the whole label classification task. The results in Table 2 shows no significant difference on prediction results between the divisions while data amount over about 100 records, except of Department of Traumatology and Department of Dermatology. The diagnosis in these two departments partially depend on visual inspection, thus, leading to incomplete expression in text information.

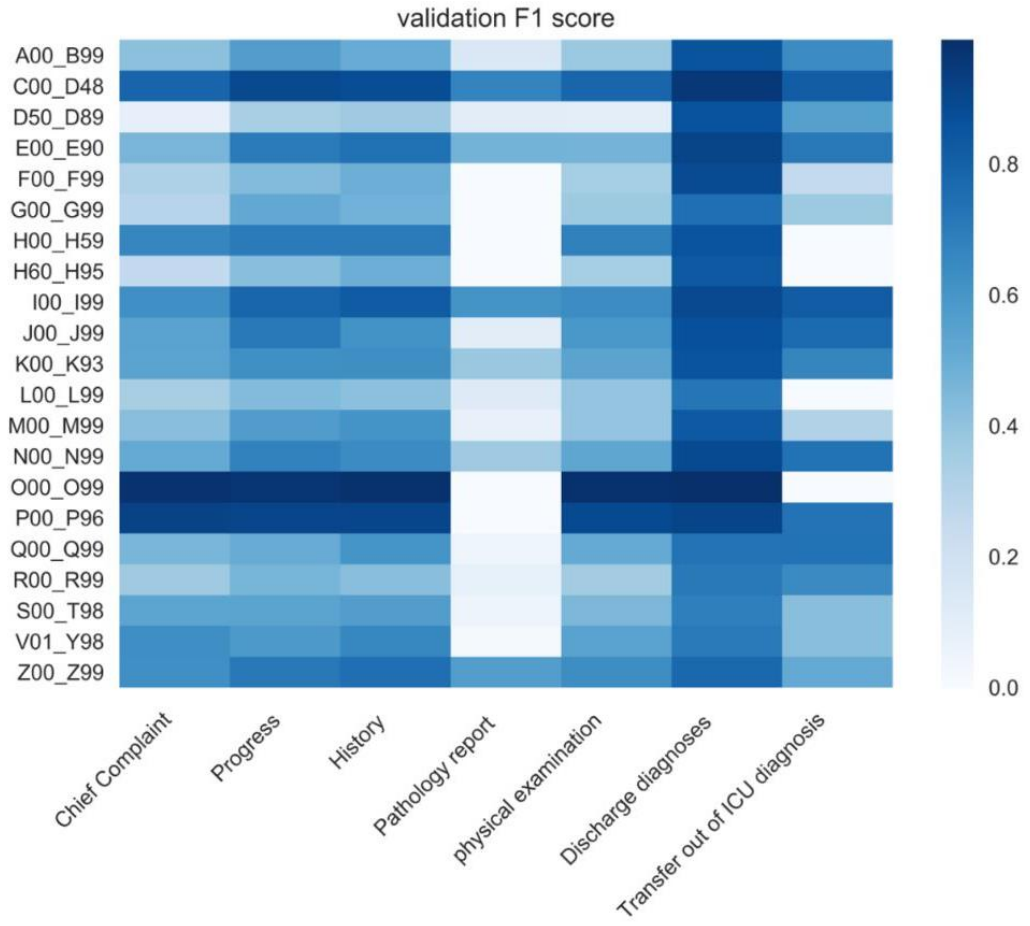


Fig. 2: Performance comparison with different input free-text data using F1 score in validation datasets.

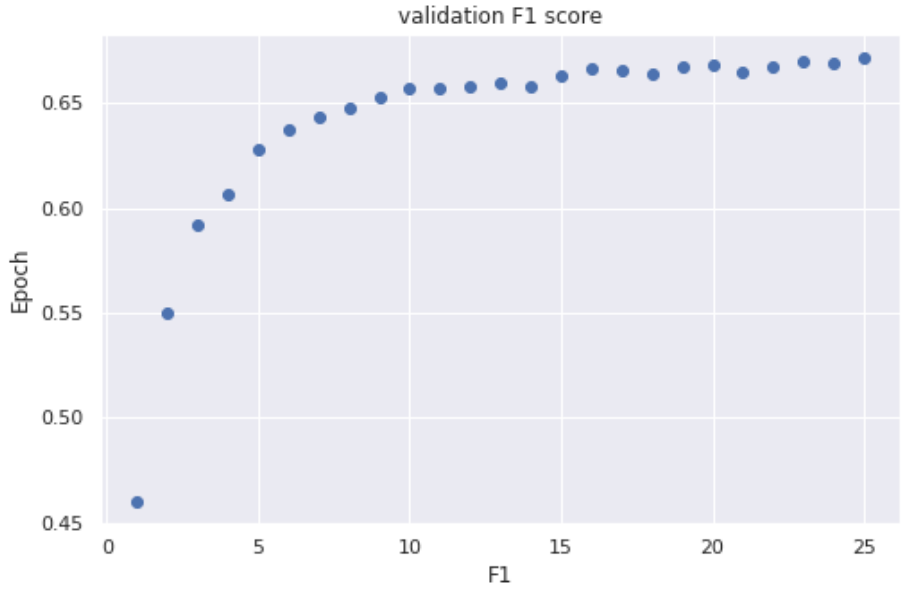


Fig. 3: F1-score performance on CM whole label prediction in validation dataset.

3.3. ICD-10 PCS Whole Label Classification

In ICD-10-PCS whole label classification task, the complete ICD-10-PCS code, i.e. PCS codes with 7 characters, corresponding to discharge diagnosis records are 9,513 labels. Progress, discharge diagnosis and physical examination records are applied for training DNN model. Result in Figure 4 implies that our model can achieve 0.58 on f1-score and 0.81 on recall@20 with progress as input data and word embedding size of 300 dimension.

Table 2: Validation data amount and the prediction for each departments on CM codes.

Department	F1-score	Data amount
Pediatrics	0.671	1,809
Orthopedics Surgery	0.68	1,554
Oncology	0.661	1,647
Obstetrics & Gynecology	0.67	2,136
Dentistry	0.668	258
Internal Medicine	0.669	5,470
Urology	0.67	1,532
Traumatology	0.678	293
Otolaryngology	0.667	929
Surgery	0.673	5,996
Neurology	0.663	243
Ophthalmology	0.674	631
Physical Medicine & Rehabilitation	0.659	104
Emergency Medicine	0.672	76
Family Medicine	0.681	256
Psychiatry	0.693	192
Dermatology	0.649	129
Geriatrics & Gerontology	0.654	6

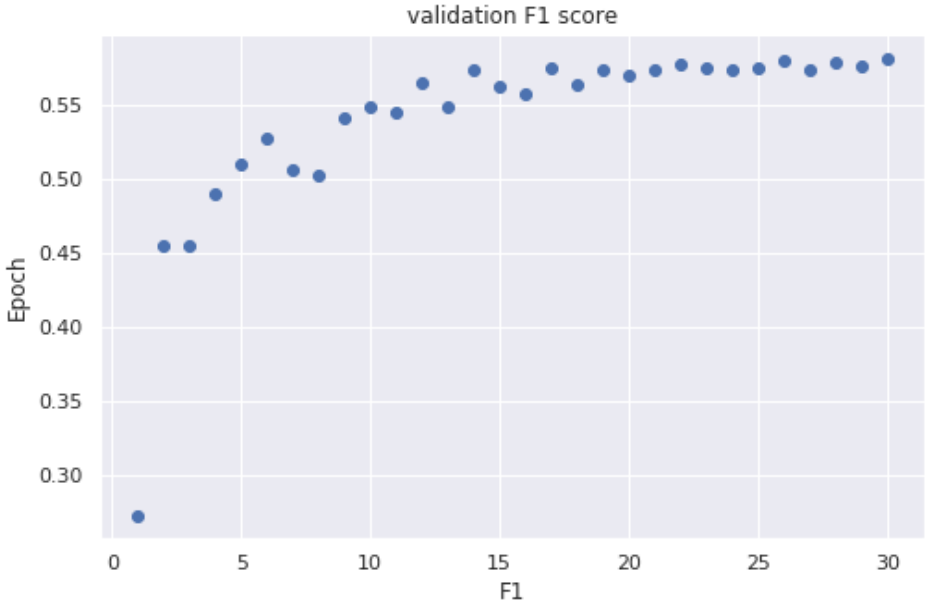


Fig. 4: F1-score performance on PCS whole label prediction in validation dataset.

3.4. ICD-10 Auto-coding System

The target of this research is to build an ICD-10 auto-coding system for assisting disease coders to elevate the work efficiency and coding accuracy. An ICD-10 auto predicting interface by taking discharge diagnosis as reference is published on http://nets.csie.ntu.edu.tw for accelerating coding efficiency. Architecture of the predicting system is shown in Figure 5. DNN model executed by python script will return the top 20 highest ICD-10-CM and ICD-10-PCS codes with recall@20 of 0.87 and 0.81 separately. The predicting process of each case takes less than 30 seconds which dramatically shorten the coding time of 30 mins per case in average.

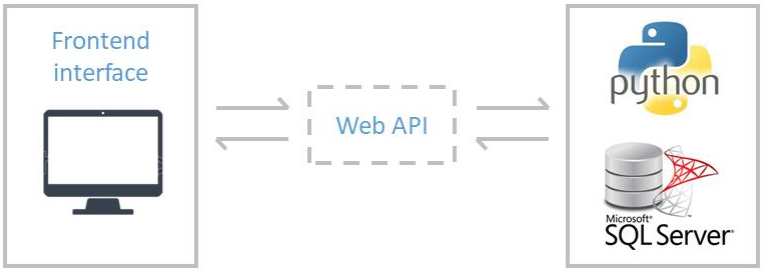


Fig. 5: ICD-10 auto-coding interface pipeline.

4. Conclusion

An ICD-10 classification model developed by NLP and deep learning model without any background knowledge from EHR data is realized in our research with f1-score 0.67 and 0.60 in CM and PCS, respectively. Also, the well-trained model is applied to the web service for assisting disease coders on ICD-10 coding work. In the future, we will keep working on applying BioBERT embedding approach and building a rule-based system to improve the ICD-10-CM classification task.

5. Acknowledgements

This study was supported by grants from the Ministry of Science and Technology, Taiwan (MOST 107-2634-F-002-015). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The authors have declared that no competing interests exist.

6. References

[1] WHO. 2017. ICD-10 Version: 2015. apps.who.int. (May 2017).

[2] Mills, Ronald E. “Estimating the impact of the transition to ICD-10 on Medicare inpatient hospital payments.” ICD-10 Coordination and Maintenance Committee presentation, March 15, 2015, Baltimore, MD.

[3] R. Caruana and A. Niculescu-Mizil, “An empirical comparison of supervised learning algorithms,” in Proceedings of the 23rd International Conference on Machine Learning (ICML 2006), 2006, pp. 161–168.

[4] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In Advances in Neural Information Processing Systems, pages 3111–3119, 2013.

[5] KyungHyun Cho Junyoung Chung, Caglar Gulcehre and Yoshua Bengio. 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. arXiv:1412.3555[cs] (Dec. 2014).

[6] Jurgen Schmidhuber Felix A. Gers. 2001. LSTM recurrent networks learn simple context free and context sensitive languages. IEEE Transactions on Neural Networks 12, 6 (2001), 1333–1340.

[7] Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Liu PJ, et al. Scalable and accurate deep learning for electronic health records. 2018.

ISBN 978-981-14-4787-7
Proceedings of 2020 the 10th International Workshop on Computer Science and Engineering (WCSE 2020)
Yangon (Rangoon), Myanmar (Burma), February 26- February 28, 2020, pp. 52-56

Innovation Security of Beaufort Cipher by Stream Cipher Using Myanmar-Vigenere Table and Unicode Table

Htet Htet Naing[1+], Zin May Aye [2]

[1] University of Computer Studied (Yangon), Myanmar

[2] University of Computer Studied (Yangon), Myanmar

Abstract. Nowadays, securing information and message transformation are going with electronic way, the security becomes very important role on public network. Cryptography is readable message convert to unreadable message using encryption/decryption process. Encryption Process is sender and decryption process are receiver side. Commonly, information can be storing with international language such as English language. At the present time, everyone is trying to be more secure not only English but also own language such as Myanmar, Chinese, Tamil etc. Confidential data are transferred through with regional language by using with more innovative method. To secure such information, encryption/decryption plays an important role in information security. In cryptography, there are several cipher techniques such as, polyalphabetic cipher, Stream cipher, Block cipher etc. This section using Beaufort cipher is an example of substitution cipher, In this paper, we propose an advanced encryption algorithm that improves the security of Beaufort encryption by combining it with a modern encryption method such as Stream cipher for the Myanmar language, Stream cipher is considered relatively as an unbreakable method and uses a binary form (instead of characters) where Plain text, encrypted text and key are bit string.

Keywords: cryptography, encryption, decryption, Beaufort Cipher, Stream Cipher

1. Introduction

Symmetric and Asymmetric are the two types of encryption. In symmetric encryption techniques we use the same key for both encryption and decryption purpose [1]. Asymmetric-key encryption using public and private keys, the public key is announced to all members while the private key is kept secure by the user. The sender uses the public key of the receiver to encrypt the message. The receiver uses his own private key to decrypt the message. In symmetric method, there are two techniques (substitution and transposition) are used as a classical method [1]. The Beaufort cipher, is a substitution cipher similar to the Vigenère cipher, with a slightly modified enciphering mechanism and tableau [2]. Its most famous application was in a rotor-based cipher machine. Substitution has further two types, Mon oalphabetic and polyalphabetic cipher [3]. In monoalphabetic the character in the Plaintext is changed to the same character in the Ciphertext. In polyalphabetic cipher a single character in the Plaintext is changed to many characters in the Ciphertext .Permutation technique is one in which the Plaintext remains the same, but the order of characters is shuffled around to get the Ciphertext .Also the symmetric ciphers can be divided into Stream ciphers and block ciphers, as a modern ciphers [4]. Stream ciphers encrypt the digits (typically bytes), or letters (in substitution ciphers) of a message one at a time. Block ciphers take a number of bits and encrypt them as a single unit, padding the plaintext so that it is a multiple of the block size [5].

2. Background Theory

2.1. Vigenere Cipher

[+] Corresponding author. Tel.: +959428312012.
E-mail address: htethtetnaing@ucsy.edu.mm

The Vigenere cipher is a plain-text form of encoding that uses alphabetical substitution to encode text. The Vigenere cipher, like other contemporary cryptographic ciphers, uses something called a tabula recta .The encryption of the original text is done using the *Vigenère square or Vigenère table.*

- The first row of this table has the 26 Character. Starting with the second row, each row has the letters shifted to the left one position in a cyclic way. The table consists of the alphabets written out 26 times in different rows, each alphabet shifted cyclically to the left compared to the previous alphabet, corresponding to the 26possible
- The first row of this table has the 26 Character. Starting with the second row, each row has the letters shifted to the left one position in a cyclic way. The table consists of the alphabets written out 26 times in different rows, each alphabet shifted cyclically to the left compared to the previous alphabet, corresponding to the 26possible.
- The alphabet used at each point depends on a repeating keyword.

Table 1: Vigenere Table

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
B	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A
C	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B
D	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C
E	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D
F	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E
G	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F
H	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G
I	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H
J	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I
K	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J
L	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K
M	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L
N	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M
O	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N
P	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
Q	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
R	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
S	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
T	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
U	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
V	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
W	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
X	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
Y	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
Z	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y

2.2. Beaufort Cipher

Beaufort cipher is a polyalphabetic substitution cipher and a variant of Vigenère cipher. Encryption and decryption using Beaufort cipher is achieved though the same algorithm. To encrypt a message, repeat the Keyword above the Cipher. If the plaintext is “S” with key “T”. Find the column with “S” on the top and travel down that column to find key “T”. Travel to the left edge of the tableau to fine the cipher text. To decrypt, the reverse the encryption process. The Beaufort Cipher using the Vigenere Cipher Table.

The Beaufort cipher can be described algebraically. The Beaufort cipher using an encoding of the letters A-Z as the numbers 0-25 and using addition modulo 26, let M=M1……Mn be the characters of the message, C= C1….Cn be the characters of the cipher text and K=K1….Kn be the character of the key, repeated if necessary. Then Beaufort encryption E is written,

$$C=EK(Mi) = (Ki- Mi) \bmod 26$$

Similarly, decryption D using the key K

$$Mi=Dk(Ci) = (Ki -Ci) \bmod 26$$

2.3. Stream Cipher

A stream cipher is a symmetric key cipher where plaintext digits are combined with a pseudorandom cipher digit stream (keystream). In a stream cipher, each plaintext digit is encrypted one at a time with the corresponding digit of the keystream, to give a digit of the ciphertext stream. Since encryption of each digit is dependent on the current state of the cipher, it is also known as **state cipher**. In practice, a digit is typically a bit and the combining operation is an exclusive-or (XOR) [6].

The Stream cipher is classified as a *synchronous* stream cipher. By contrast, *self-synchronizing* stream ciphers update their state based on previous ciphertext digits [7].

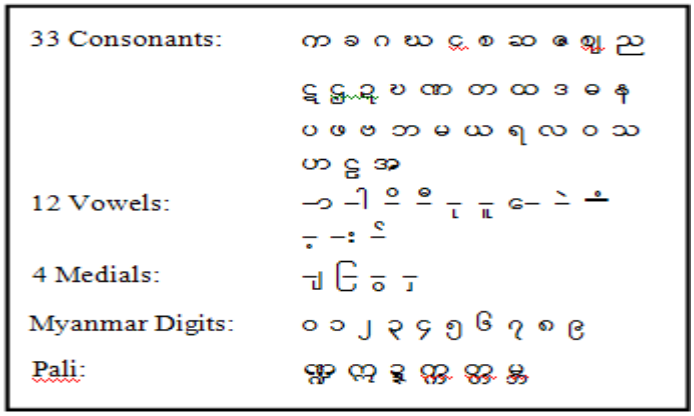


Fig. 1: Sample a set of Myanmar Alphabet

3. Proposed System

In normal Vigenere table uses a 26×26 table with A to Z as the row heading and column heading. Traditionally Myanmar Fonts have been pretending to be English fonts. This means that they are using the numbers allocated for the Latin alphabet to represent Myanmar characters. Our paper propose method uses a 31×31 table in Myanmar Language Vigenere table. These 31 characters is the mostly used of Myanmar character. This table is usually referred to as the Original Vigenère Table. The first row of this table has the 31 Myanmar Character. Starting with the second row, each row has the letters shifted to the left one position in a cyclic way. The table consists of the alphabets written out 31 times in different rows, each alphabet shifted cyclically to the left compared to the previous alphabet, corresponding to the 31 possible. Modern cipher is normally use combination of substitution with transposition. In this paper, we proposed a new combination method Beaufort cipher by using Myanmar Vigenere Table. In this paper, Myanmar Character □-□ as the number 0-31 after getting the cipher text. In Stream Cipher, half of work, for example the user enters the message in Myanmar language and this is divided into each work. And then half of work the message length. For example, the message length is 20, the first 10 Myanmar character are encrypted with Beaufort Cipher and last 10 Myanmar character are encrypted with Stream Cipher. The stream cipher process using binary code, so Myanmar Character Unicode are converted to Binary code [7-8].

Table 2: Propose Myanmar Vigenere Table

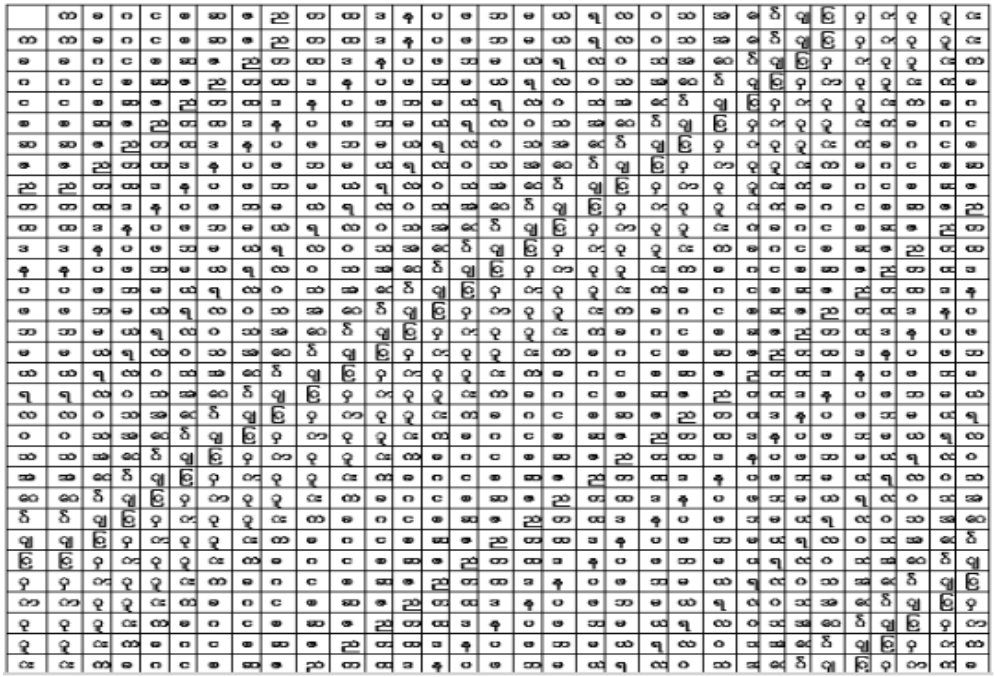


Table 3: Sample of Myanmar Unicode And Binary Code Table

က 10000000	ခ 10000001	ဂ 10000010	ဃ 10000100	စ 10000101	ဆ 10000110
ဇ 10000111	ည 10001010	တ 10010000	ထ 10010001	ဒ 10010010	န 10010100
ပ 10010101	ဖ 10010110	ဘ 10011000	မ 10011001	ယ 10011010	ရ 10011011
လ 10011100	ဝ 10011101	သ 10011110	အ 10100001	ဧ 10110001	ဩ 10110101
ျ 10111011	ြ 10111100	ွ 10111101	ာ 10101100	ု 10101111	ူ 10110000
း 10111000					

3.1. Step by Step of Proposed System Process

1. Start
2. Read Plaintext P
3. Read Key K
4. Using Beaufort equation C =K-M or Beaufort cipher by using Vigenere Table the characters in First half locations of Plaintext.
5. Apply stream cipher to encipher each character in the second half location as follows:
   - Converting the characters to Unicode value then to equivalent binary form.
   - Enciphering these characters using stream cipher equation C = P $\oplus K_{bin}$.
   - Converting the resulted binary numbers to equivalent Unicode value then to characters to obtain the Cipher characters.

Plaintext: ကျောင်းသားများ စာဖတ်နေသည်

(English : the students are learning)

Key: စာအုပ်

(English : Book)

Plaintext: ကေ ကျ တာ c ်ိ း သ တာ း မ ျ တာ း စ တာ ဖ တ ်ိ ေ န သ ည ်ိ

Key: စ တာ အ ု ပ ်ိ စ တာ အ ု ပ ်ိ စ တာ အ ု ပ ်ိ စ တာ အ ု

	0	1	2	3	4	5	6	7	8	9	10	11
Plain text	ေ	က	ျ	တ	c	်ိ	း	သ	တ	း	မ	ျ
Key	စ	တ	အ	ု	ပ	်ိ	စ	တ	အ	ု	ပ	်ိ
Cipher Text	ဖ	တ	ု	ခ	ထ	က	ဆ	ည	ြ	ူ	ု	း

Table 4: Encryption Beaufort Cipher of First Half Location by Using Myanmar-Vigenere Table

	13	14	15	16	17	18
Plain text	တ	း	စ	တ	ဖ	တ
Unicode equivalent	4140	4152	4101	4140	4118	4112
Key	စ	တ	အ	ု	ပ	်ိ
P_{bin} $\ominus$ K_{Bin}	10101100 10000101	10111000 10101100	10000101 10100001	10101100 10101111	10010110 10010101	10010000 10110101
C_{Bin}	00101001	00010100	00100100	00000011	00000011	00100101
CipherText	ဩ	န	ဤ	ဃ	ဃ	ဥ

Table 5: Encryption of Stream Cipher for Second Half Location by Using Myanmar Unicode Table

$C=P_{bin} \oplus K_{bin}$

	19	20	21	22	23	24
Plain text	ိ	ေ	ာ	သ	ည	ိ
Unicode equivalent	4149	4145	4116	4126	4106	4149
Key	ဓ	ဝ	ဆ	ဍ	ဂ	ဒ
P_{bin} ⊕ K_{Bin}	10110101 10000101	10110001 10101100	10010100 10100001	10011110 10101111	10001010 10010101	10110101 10110101
C_{Bin}	00110000	00011101	00110101	00110001	00011111	00000000
CipherText	ွ	ံ	ိ	ေ	ဟ	ဝ

4. Scope and Limitation of Proposed System

Proposed System aimed to the security of Myanmar Language. Security in cryptography is based on how secure the algorithm is against various attacks for Myanmar Language. Myanmar Language contain consonants, vowels, Medial, virama, Myanmar digits and Pali. Some Limitation of our proposed system. In our system can use only Myanmar Characters on the above table (Table 2). I didn't think about Pali (ပါဠိ) for Myanmar Language.

5. Conclusion

Beaufort cipher regard as simplest and weakest method, that mean it is very easy to attack. To overcome the limitations of this method, we propose a new algorithm which includes combining Beaufort substitution cipher with Stream cipher. We notice that repeated portions of plaintext always encrypted with the different portion of the keyword or binary key, because we encipher the letters in first location with Beaufort cipher and the letters in second locations with Stream cipher, result in different ciphertext segments, that mean proposed algorithm hides the relationship between Ciphertext and Plaintext, and makes the cryptanalysis more difficult. Furthermore, the proposed combination method enhances the security of Vigenere method and make the detection process not easy, because the Stream cipher relatively regards as unbreakable. This paper attempts to enhance the encryption / decryption of the regional language.

6. References

[1] Paar C. and Pelzl J. 2010, Understanding Cryptography, Springer-Verlag Berlin Heidelberg.

[2] https://en.wikipedia.org/wiki/Beaufort_cipher

[3] Fairouz Mushtaq Sher Ali, Falah Hassan Sarhan "Enhancing Security of Vigenere Cipher by Stream Cipher". International Journal of Computer Applications (0975 –8887). Volume 100–No.1, August2014

[4] T.M. Aung, H.H. Naing"AComplexTransformationofMonoalphabeticCiphertoPolyalphabeticCipher:(Vigenère-AffineCipher)". International Journal of Machine Learning and Computing, Vol. 9, No. 3, June 2019

[5] https://en.wikipedia.org/wiki/Block_cipher

[6] https://www.Utf8-chartable.de/Unicode-utf8-table.pl?start=4096&number=128&utf8=string-literal

[7] https://en.wikipedia.org/wiki/Stream_cipher

[8] "https://www.unicode.org/notes/tn11/UTN11_3.pdf

ISBN 978-981-14-4787-7
Proceedings of 2020 the 10th International Workshop on Computer Science and Engineering (WCSE 2020)
Yangon (Rangoon), Myanmar (Burma), February 26- February 28, 2020, pp. 57-62

Development of Spoken Language Recognition System for Humanoid Robot

Khaing Yee Mone [1+], Yoshio Yamamoto [2]

[1] Course of Mechanical Engineering, Graduate School of Engineering, Tokai University, Hiratsuka, Japan
[2] Department of Precision Engineering, School of Engineering, Tokai University, Hiratsuka, Japan

Abstract. "Pepper" robot has a variety of embedded abilities of interaction to communicate with humans such as Speech, Image, Detection, Gesture, and Tablet Service. Our research focused on interaction by speech recognition system of Pepper robot and analyzed its effectiveness. Since there are some limitations in built-in speech recognition system, we combine it with cloud-based speech recognition. The purpose of this research is to get the correct and reliable recognition of verbal speeches from users. Therefore, Speech API from Google Cloud service was used to implement more interactive behavior for Pepper robot.

Keywords: Pepper Robot, Human Robot Interaction, Cloud-based Speech Recognition

1. Introduction

Humanoid robots are increasingly appearing in daily contexts of people's lives, for example, they can assist humans in customer service, welcoming, informing and amusing humans in a kind and familiar manner. The deployment of robots in all these potential applications enriches our lives and, robot's social capability and widespread intelligence are needed to improve more than ever. In this regard, human–robot interaction (HRI) have been drawing considerable attentions in the robotic research community and substantial amount of efforts have been devoted to humanoid robots to improve robot's social skill [1].

Pepper robot, released from Softbank Robotics, is very popular in these kinds of human robot interaction applications. Initially it was particularly designed for an application of business uses in Softbank stores, it becomes a platform of interest for various other applications including academics and home-entertainment areas. Pepper robot was designed by these principles: pleasant appearance, safety, interactivity, affordability and good autonomy [2]. Pepper robot has a smart appearance with its humanoid upper body, operated with amazing functionalities such as emotion perception, speaking with gestures and omnidirectional movement provided by the wheeled base. It can interact with humans through speech, and if the customer requests some information, it can also give answer by either speech feedback, or visual feedback using a tablet [1].

Fig. 1: Waving posture of Pepper robot.

[+] Corresponding author. Tel.: + 81 80 9566 7021
E-mail address: 8bemm084@mail.u-tokai.ac.jp

Fig. 1 shows the waving posture of Pepper robot, welcoming students in our laboratory. Additionally, Pepper's face recognition function is for learning about the users, and continuously updating its knowledge base about the users. For extended interaction capabilities, the android tablet on Pepper's chest can be used either for developing apps that integrate with the robot or as a display by loading web pages, pictures or videos [2]. Although most features on Pepper robot are pre-programmed and demonstrate a conscious thinking process, its structural design and software capabilities offer a great predisposition for human interaction [3]. Therefore, researchers are trying to improve them in order to provide more personalized human-robot interactions.

The aim of our research is to create a behavior for Pepper in which Pepper can imitate human's speech. For this purpose, we investigated the possibility of improvement of speech recognition by using cloud-based speech recognition. It enables robots to access large amount of computing power and Pepper's application can interrogate databases and formulate the most appropriate response.

This paper is organized as follows: we present the experience from the operation of Pepper humanoid robot and its software environment in the next section, these are very important to understand for creation of new robot behaviors, and then we show the development of speech recognition system and the behavior consideration and implementation. In the last section, we discuss current results and describe our future work.

2. Pepper for Human Robot Interaction

Pepper is a fully autonomous humanoid robot designed by Aldebaran Robotics, and released in 2015 by SoftBank Robotics (SoftBank acquired Aldebaran Robotics in 2015). In 2015 Pepper was available only in Japan, in June 2016 also in Europe, and in November 2016 in the US [4]. Pepper robot seems to be one of the best options for implementing and research on HRI, and its height and physical proportions and dimensions appearing more like a human-being and resulting better interaction in HRI. The hardware design and functionality are preprogrammed in a form of an API (application programming interface) [5]. A significant advantage of this robot design is its full programmability.

2.1. Embedded Software

NAOqi is the name of the main software that runs on the robot and controls it. NAOqi provides programming framework to develop applications on the Softbank's robots: NAO and Pepper. Different software development kits are provided for any of these programming languages: Python, C++, Java, and Robot Operation System (ROS).

The NAOqi process provides lookup services to find methods, and network access, allowing methods to be called and executed remotely. Local modules are in the same process, so they can share variables and call each other methods without serialization nor networking, thereby allowing the fast communication. Local modules are ideal for closed loop control. Remote modules communicate using the network, and so it is impossible to do fast access using them [6]. Fig.2 shows the process of NAOqi framework.

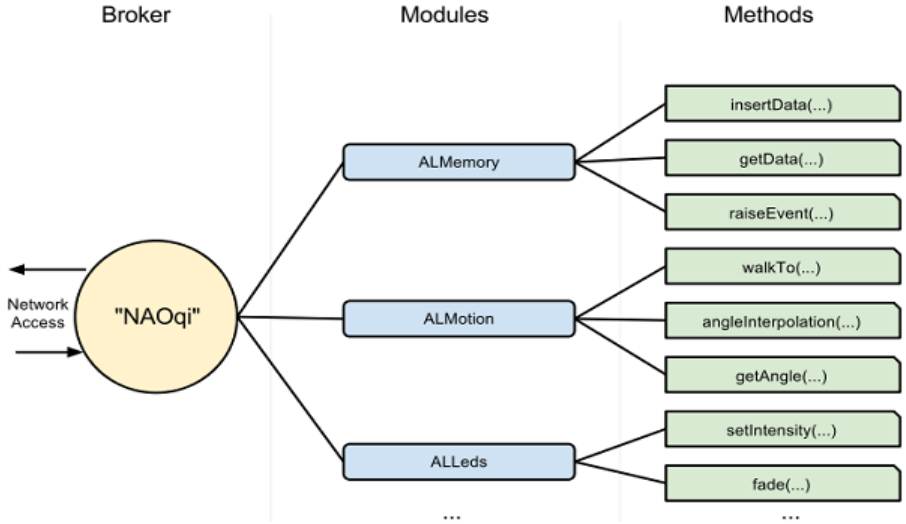


Fig. 2: NAOqi Operating System [6].

The NAOqi executable comes with a list of core modules and a public APIs, which are functionally divided in groups. The default APIs are: NAOqi Core, NAOqi Motion, NAOqi Audio, NAOqi Vision, NAOqi People Perception, and NAOqi Sensors. Among them, the two important modules, used in our interaction behavior creation, are NAOqi Audio and NAOqi People Perception.

NAOqi Framework provides Pepper robot to have an autonomous life, having the basic awareness capabilities and seemingly alive. This shows that Pepper robot is different from others, simple, active and ready to help or interact with humans. And what is the most interesting factor for developers is to develop more interactive behaviors.

Since Pepper has been popular in media as a conversational agent, the work presented in [7] used IBM Bluemix Speech Recognition Service to enhance the robot ability to interact with its users. They integrated Pepper's NAOqi software with ROS to improve the autonomy and making Pepper robot moves autonomously in an environment with humans and obstacles.

Previous work [8] showed that their integration of Pepper with state-of-the-art vision and speech recognition system, and they introduced a learning algorithm to improve communication capabilities over time, which can update speech recognition through social interaction.

2.2. Experience with Pepper Humanoid Robot

This section shows user's experiences with desktop software to control Pepper remotely. One of the effective software to control Pepper is "Choregraphe". The Choregraphe environment is supposed to be used for constructing an application by using some of the built-in function blocks/objects; that are grouped into thematic libraries [9]. It becomes possible to modify functionality of a block/object by developing its Python code. These blocks/objects can be easily and intuitively used for building custom applications with varying degrees of complexity [10]. By working Choregraphe with Python SDK, users can have full access to all build-in modules and can expand robot's functionality with new, practically unlimited resources.

It is a powerful graphical environment that allows the user and developer to work with Pepper with a friendly interface and easy to understand behavior creation, we use Choregraphe version 2.5.5.5. We have the following experiences by using Choregraphe;

- Creation of animations, behaviors and dialogs,
- Testing them on a simulated robot, or directly on a real one,
- Developing Choregraphe behaviors with Python code.

We had successfully done the presentation behavior for Pepper robot created in Choregraphe for the purpose of helping teachers. We presented our work at JSME Robotics and Mechatronics Conference, 2019 in Hiroshima [11].

3. Development of Speech Recognition

In this section, we describe the approach to develop Pepper's speech recognition and people perception. We studied the robustness of Pepper's built-in speech recognition and combination of the existing system with cloud-based speech recognition to improve the accuracy of Pepper's speech interaction.

3.1. Built-in Speech Recognition

Pepper speech recognition system runs by Nuance Solution, a compact speech solution for embedded systems, which is accessed through NAOqi, Speech-To-Text module. This module process speech recognition function with the help of four microphones, placed in the head to provide sound localization, and two loudspeakers, laterally placed on the left and the right sides of the head [12].

However, there are two key limitations for the speech recognition module. The first limitation is insufficient variety of languages since Nuance solution offers only a few languages to choose from, and furthermore the embedded library can recognize only phrases from a predefined set. The human users need to set the vocabulary of phrases to be recognized. Whenever the input audio is matched with a phrase inside the vocabulary, Pepper robot can recognize the audio and understand human's speech. Recognition accuracy is considerably lower for a longer sentence, as the software has more opportunities to make errors. When users say something that is not included in the vocabulary, sometimes, the built-in speech recognition

software is not enough and wrongly matches it to a phrase in the vocabulary. The second one is speech signal weakness. During the conversation, the movement of the robot's head mechanisms (motors, fan) introduces interference. That can slightly change in the strength of the speech signal, which can reduce the quality of speech recognition.

Therefore, the sound collected from Pepper's microphones require to be filtered to remove the noise affected by head's movement. Furthermore, we would like Pepper to learn about its environment and how humans prefer to interact with it, and for that we need Pepper to be able to recognize previously undefined speech.

3.2. Cloud Speech API

In order to solve the limitations encountered in Pepper's speech recognition, we used cloud speech API (speech-to-text service) provided from Google. Google Cloud Speech is a cloud-based streaming speech recognition software platform. Google's speech algorithm consists of a deep neural network that has been trained on a large amount and variety of speech from Google users, and is able to recognize general speech [13]. Users connect to the service and send streaming audio, and the service returns transcription results in real-time, along with a confidence level. While Pepper works with cloud-based speech recognition, the creation of robot's applications can be extended as Natural Language Processing, language translation and free word recognition as shown in Fig. 3.

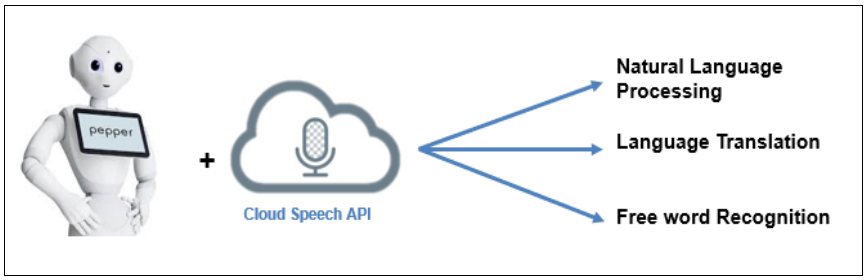


Fig. 3: Pepper robot working with cloud speech API

4. Knowledge Obtained from Trial Results

We tried several experiments for cloud-based speech recognition with the purpose to design minimum hardware requirements and to get the correct operation of speech recognition service. Unlike Pepper's speech recognition, users do not specify the speech they expect to receive. The input audio from each of Pepper's four-microphones is streamed to Google and use the recognition result with the highest confidence. The cloud-based speech recognition service allows for more general speech and appears to perform somewhat better, but it also requires an active internet connection with enough speed for streaming audio.

Table.1 shows the comparison of recognition accuracy between built-in NAOqi speech API and cloud speech API from our trial results. Our experiments work on Pepper robot with NAOqi version 2.9, and we used three kinds of speeches; simple speech (one word), short sentence (three words) and random sentence (more than five words). These results come out from fifty trial times by one user. It shows that using Google cloud speech has higher accuracy than built-in system. And we consider that the possibility of combination of two systems can get the highest accuracy.

Table 1: Comparison of recognition accuracy

Speech Recognition Software	Recognition Accuracy
NAOqi Speech-To-Text	0.56
Google cloud speech API	0.70

The important factor of this research is to test the sound transmission time to the Google Cloud service and resultant text transmission time to robot. We expected that the acceptable maximum processing time between the ends of the human speaking to the beginning of the robot speaking is 2 seconds. During this time, the system will perform the following steps;

• Perform Speech-to-Text on Google cloud service

• Return transcriptions to the robot control system

• Initialize Text-to-Speech module to make robot speak

At the present stage, the speech-to-text result is received immediately after sending human speech to Google cloud. It shows that these steps can be performed in total time of less than 2 seconds. The following section describes the idea of interaction behaviour for Pepper robot.

4.1. Interaction Behaviour

The Pepper robot is supported to be great for entertainment or amusement purposes. We would like to make Pepper imitate human speech after human says something. This means that Pepper robot hears human speaking and speech recognition will be done by cloud speech and then the speech-to-text transcription will be sent to Pepper tablet immediately. And robot's control system will connect the tablet and NAOqi to pronounce the text on the tablet. We show the general configuration of our system in Fig. 4.

Fig. 4: Configuration of our interaction behaviour

5. Summary

This paper presented the idea to develop human's spoken speech recognition of Pepper robot. The purpose is to improve the familiarity between robot and human through communication. At the present stage, we have done with trial experiments to test the robustness of Google cloud speech API and the strength of audio signals recorded by Pepper's microphones. The speech-to-text response time form Google cloud is at acceptable level and it exhibits better accuracy if we can reduce noise affected on audio signal. For the ongoing work, we would like to conduct on behaviour creation that Pepper can imitate human speech.

6. Acknowledgements

We would like to offer special thanks to JICA Innovative Asia Program for their funding support for this research and particularly grateful for opportunities of study in Japan by JICA Innovative Asia Program.

7. References

[1] IEEE Spectrum. How Aldebaran Robotics Builds Its Friendly Humanoid Robot, Pepper. Retrieved February 2015, from https://spectrum.ieee.org/robotics/home-robots/how-aldebaran-robotics-builds-its-friendly-humanoid-robot-pepper, 2014.

[2] SoftBank Robotics, "SoftBank Guidelines Documentation", available on-line (2018-04-01): http://doc.aldebaran.com/download/Pepper_B2BD_guidelines

[3] A. Gardecki, M. Podpora, "Experience from the operation of the Pepper humanoid robots", 978-1-5386-1528-7/17/$31.00, 2017 IEEE.

[4] N. M. Hombur "Designing HRI Experiments with Humanoid Robots: A Multistep Approach", In *Proc. of the 51st Hawaii International Conference on System Sciences*, 2018.

[5] Pepper Robot Programming tutorials from, https://www.about_robots.com/pepper_robot_programm ing.html

[6] *NAOqi APIs and Documentations*, http://doc.aldebaran.com/2-5/naoqi/index.html/

[7] V. Perera, T. Pereira, J. ConnIell, and M. M. Veloso, "Setting up pepper for autonomous navigation and personalized interaction with users," *In CoRR*, vol. abs/1704.04797, 2017. [Online]. Available: http://arxiv.org/abs/1704.04797

[8] Michiel de Jong, Kevin Zhang, Aaron M. Roth, Travers Rhodes, Robin Schmucker, Chenghui Zhou, Sofia Ferreira, Jo ão Cartucho, and Manuela Veloso. "Towards a Robust Interactive and Learning Social Robot". In *Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS2018),* Stockholm, Sweden, July10–15, 2018, IFAAMAS, 9pages

[9] Choregraphe Tutorials, “how to script the python boxes” http://doc.aldebaran.com/2-1/software/choregraphe/tutos/

[10] E.Pot, J.Monceaux, R.Gelin, and B.Maisonnier, “Choregraphe: a Graphical Tool for Humanoid Robot Programming,” Aldebaran Robotics, ResearchGate, 2009.

[11] Khaing Yee Mone “Advanced Presentation Behavior for Pepper Robot Based on Speech Recognition”. In *proceedings of the Robomech, Robotics and Mechatronics Lecture 2019 in Hiroshima*, June 2019.

[12] G. Suddrey, A. Jacobson and B. Ward, “Enabling a Pepper Robot to provide Automated and Interactive Tours of a Robotics Laboratory”, In *the International Conference on Intelligence Robots*, 2018 IEEE.

[13] Google Cloud Speech, from https://cloud.google.com/speech/

ISBN 978-981-14-4787-7
Proceedings of 2020 the 10th International Workshop on Computer Science and Engineering (WCSE 2020)
Yangon (Rangoon), Myanmar (Burma), February 26- February 28, 2020, pp. 63-67

Building Large Scale Text Corpus for Joint Word Segmentation and Part-of-Speech Tagging of Myanmar Language

Dim Lam Cing [1+], Khin Mar Soe [2]

[1,2]Natural Language Processing Lab (NLP), University of Computer Studies, Yangon, Myanmar

Abstract. In Natural Language Processing (NLP), Word segmentation and Part-of-Speech (POS) tagging are fundamental tasks. The POS information is also necessary in NLP's preprocessing work applications such as machine translation (MT), information retrieval (IR), etc. Currently, there are many research efforts in word segmentation and POS tagging developed separately with different methods to get high performance and accuracy. Word segmentation and Part-of-speech tagging is one of the important actions in language processing. Against this, while numerous models are provided in different languages, few works have been performed for Myanmar language. This paper describes the building of Myanmar Corpus to use for joint word segmentation and part-of-speech tagging of Myanmar Language. In our research, the corpus contains 51207 sentences and 839161words. The corpus is created using 12 tags. To evaluate the accuracy of the corpus, HMM model is trained on different data size and testing is done with closed test and opened test. Results with 94% accuracy in the experiments show the appropriate efficiency of the built corpus.

Keywords: Natural Language Processing, POS, HMM, Corpus

1. Introduction

Language corpora are widely used in linguistic research and language technology. A tremendous interest has arisen in recent years in building and developing computerized language corporations. Studying the electronic corpora of different languages provides learners and researchers with the opportunity to work with language information with a variety of tools and techniques in analytical procedures and programs [1].

POS tagged corpus is a structured textual database that serves as a reference material for further NLP work as well as a learning repository for machine translation algorithms and other applications for code. Building syntactically classified corpus requires a sequence of procedures such as text preprocessing, tokenizing sentences and POS tagging. Also it is influenced on all areas of NLP such as information retrieval, text-to-speech, parsing, information extraction and any linguistic research for corpora [2].

While many words can be unambiguously associated with one POS or tag, other words match multiple tags, depending on the context that they appear in [3]. Therefore, the accuracy of a tagger depends on its learning database or its training data. The larger the corpus size, the better the accuracy for tagging. Also, an automatic part-of-speech tagger requires a large corpus because hand annotating is tedious task and also assigning POS tags to each word is very time consuming [4].

In this paper, we start by hand annotating raw text to build a tagged corpus. Then we process by preparing training data from the manually tagged corpus. Next, we automatically assign POS tags to each word of raw text using our proposed POS tagger. Then, we analyze result of tagged text and refine manually. We conclude with the result that POS tagged corpus for Myanmar Language is annotated by stochastic method of POS tagging.

[+] Corresponding author. Tel.: +959400423380
E-mail address: dimlamcing@ucsy.edu.mm

Myanmar Language is a common language of the national languages of Myanmar and is part of the family of the Sino-Tibetan language. It is spoken as first language by about 33 million people and as second language by 10 million people [5]. The truth is that Myanmar Language has only a small amount of linguistic computational capital. On this language, there are a few computational works. Researchers have recently started to engage in the creation and enrichment of Myanmar Language's language in the Natural Language Processing (NLP) sector. These NLP activities included the need to build a large amount of language-based corporations.

The term "corpus" is used to refer to a collection of linguistic records (masking spoken and written records) in a language for certain unique functions, and to save, take care of and translate those facts in virtual format. A corpus, as an example, can be quite small, consisting of 50,000 words or texts, or very large, such as millions of words. Corpus is the premise for linguistic research of a wide variety. The corpus range is huge. The fields of corpus-based totally studies are : grammatical research of unique linguistic production, building reference grammar, lexicography, language variation and dialectology, ancient linguistics, studies of transcription, language acquisition, language pedagogy, and processing of natural language, etc.

The need of language corpora has caused to the study of corpus linguistics. It is not a branch in linguistics, however a method that helps to carry out linguistic studies. The development of computer software program for corpus evaluation has been closely related to modern-day corpus linguistics from the very beginning. In modern corpus linguistics, linguists and computer scientists share a common goal that in order to perform any kind of linguistic analysis, it is necessary to rely on actual or real language knowledge (speech or writing). It is also an approach that addresses two main goals: how people use language in daily communication and how to create intelligent systems to communicate with people [1].

2. Related Work

To date, numerous methods of POS tagging have been presented in some languages such as English, often based on rules or statistics. But in this field there are few activities that have been done over the past few years [6]. Jabbari and Allison are doing one of the latest works on POS tagging [7]. Their strategy is based on transformation, and Brill and Hepple used it previously in English [8, 9]. Creating this tagger requires a professional learner computer that provides approximate rules. They actually applied Error-Driven Transformation Based Learning implementation. They believe their approach is 93 % accurate [10] presents an HMM model for POS tagging in Manipuri. Since Manipuri has no tagged corpus, the system uses Tagger's small set of tagged phrases based on Manipuri Rule. The system can assign tags to most of the lexical items of the test set. This tagger will be very useful in language processing applications such as text-based information retrieval, speech recognition and machine translation, etc. The proposed system can be made more efficient by applying the bigram probability to trigram probability and it gives 92 % accuracy.

There are many other POS tagging system by using Hidden Markov Model in other languages[11, 12]. The HMM method is applied in Persian POS tagging to determine the reliability of the proposed approach in simulations performed on both homogeneous and heterogeneous Persian corpus. Obtaining 98.1 % accuracy results in the experiments shows the adequate effectiveness of the Persian corpus approach proposed [11]. The research is based on a statistically based approach by measuring the probability of the tag sequence and the term likelihood of the given corpus, where the linguistic data is automatically extracted from the annotated corpus in which the tagging process is carried out. For known words, more than 90% of the accuracy can be achieved by the current tagger [12].

3. Tagged Corpus Generation

In this paper, we are mainly concerned with building the framework of Myanmar Language un-annotated raw corpus consisting of 839,161 total words and also attempting to highlight the problems faced during the construction process. A huge collection of texts would be useful for language and non-linguistic research, cross-linguistic correlations and all other communication technologies.

There are different problems related to corpus design, development and management. These issues differ depending on the corpus' form and usefulness. In fact, the development of speech corpus is different from the development of text corpus.

Myanmar language tagged corpus is essential in any applications of Natural Language Processing. There are several steps to create tagged corpus using stochastic method that are Collecting Raw Data, Manually Tagged, Preparing Training Data and Increasing Data size in the Tagged Corpus.

3.1. Collecting Raw Text

The collection of data is a vital activity to build a corpus. A great deal of raw text must be assembled from a variety of sources. Also raw text is checked for morphological and syntactic errors so as to be ready to annotation.

In case of this work, bunch of raw text are collected from online journals, newspaper and e-books. Myanmar text are copied and saved in text files.

3.2. Manually Tagged

The collecting Myanmar texts in the corpus are tagged manually by hand and have training data for statistical method.

3.3. Preparing Training Data

Currently we prepared over 50,000 sentences as training data. Then we will develop a HMM model by calculating the probabilities of the tag of each word, counting word frequency. These functions help us to analyze on tagged corpus.

3.4. Increasing Data Size in the Tagged Corpus

The corpus is enlarged by assigning POS tag automatically to unprocessed text files. POS tagger runs and assigns POS tag to each word by using the Hidden Markov Model (HMM) by selecting the maximum POS tag for each word automatically on the untagged text.

After generating tagged text, we have to analyze and refine manually to unknown tag and wrong tag. Finally, we can use these correct texts in the corpus so that our corpus size can be enlarged.

4. Experimental Evaluation

The accuracy of any Part of Speech tagger is measured in terms of the following accuracy:

$$Recall, R = \frac{Number\ of\ correct\ POS\ tag\ assigned\ by\ the\ system}{Number\ of\ words\ in\ the\ test\ set} \tag{1}$$

$$Precision, P = \frac{Number\ of\ correct\ POS\ tag\ assigned\ by\ the\ system}{Number\ of\ POS\ tag\ assigned\ by\ the\ system} \tag{2}$$

$$F_{score}, F = \frac{2PR}{P+R} \tag{3}$$

4.1. Statistic of the Dataset

For evaluation of the proposed tagger [13], a corpus having texts from different genres were used. In our corpus, consists of the Asian Language Treebank (ALT) corpus, is one part from the ALT Project and the UCSY corpus, is created by the NLP Lab, University of Computer Studies, Yangon (UCSY), Myanmar. The other data of the corpus is collected from Myanmar Grammar Book and websites that contain economic, social, art, culture, sport and religious. It aims to promote word segmentation and POS tagging research on Myanmar language. The statistic of the Dataset is described in Table 1. Although the News dataset is dominant, the data coverages the various topics.

4.2. Processing

Myanmar Language has no space between words. So, in our research, we preliminary segment the sentence in syllable using the syllable break tool [14]. And then, segment by using N-grams model (5-grams)

and select the maximum POS tag for the word by using HMM. This is implemented by using Python programming code.

Table 1: Statistic of the Dataset

Data Type	No. of Sentences	No. of Words
UCSY	5017	58301
ALT	1500	27340
Web News	35725	612875
Short Stories	1300	25375
Novels	5142	72656
Books	2523	42614
Total	51207	839161

5. Result and Discussion

To evaluate the generated corpus, two different experiments are performed. We collect 500 new sentences for closed test and open test. In our experiments, we compare four different training data size as described in Table 2. According to this experiment, the overall obtained accuracy of the training data is over 92% .

Table 2: Comparison of the accuracy on Closed Test and Open Test on different data size

Corpus Size (Total Words)	**Closed Test**			**Open Test**		
	R	**P**	**F**	**R**	**P**	**F**
547969	76%	78%	77%	69%	70%	69%
690258	81%	83%	82%	78%	79%	78%
740495	91%	92%	91%	89%	90%	89%
839161	93%	94%	93%	92%	93%	92%

Nevertheless, the results of the experiments show that the greater the training data, the greater the reliability. And there's only a little gap between the closed test and the open test in the last corpus. Finally, the accuracy of these tests in terms of accuracy show the success of building the large corpus for joint word segmentation and POS tagging for Myanmar Language.

6. Error Analysis

Some errors occurred in the experiments especially in syllable break that cannot break in some consonant appeared consecutive in the sentence. As the consequences of the syllable break error, the segmentation error occurred. The unknown words occurred because of segmentation error and person names, locations that are not containing in the corpus. The segmentation is performed by N-gram and the tagging also performed on the longest (5-grams) words, so some wrong tagging occurred. Some words of Myanmar Language have more than one POS tag. So, some POS tagging in words may cause ambiguity.

7. Conclusion

In this paper the building of Myanmar POS Corpus is described. Experimental results show that there are differences in the accuracy rate on different training data. By using a large training, joint word segmentation and the assignment of POS tagging is more accurate and reduced the unknown words, incorrect tag and ambiguous words. The paper has shown that the training corpus is efficient for joint word segmentation and POS tagging in Myanmar Language. High accuracy rate (94%) is got in closed testing of the experiment.

8. References

[1] S. Sarma, H. Bharali, A. Gogoi, R. Deka, A. Barman. A Structured Approach for Building Assamese Corpus: Insights, Applications and Challenges , *Proceedings of the 10th Workshop on Asian Language Resources*, pages 21–28, COLING 2012, Mumbai, December 2012

[2] A. Ganbold, P. Jaimai, Integrative Tools for Part-of-Speech Tagged Corpus, *School of IT, National University of Mongolia,* Ulaanbaatar, Mongolia.

[3] J. Diesner, Part of Speech Tagging for English Text Data, *School of Computer Science, Carnegie Mellon*

University, Pittsburgh.

[4] F. M. Hasan, N.UzZaman , M. Khan, Comparison of Unigram, Bigram, HMM and Brill's POS Tagging Approaches for some South Asian Languages, *Proc. Conference on Language and Technology (CLT07),* Pakistan, 2007

[5] https://en.wikipedia.org/wiki/Burmese_language

[6] M. Mohseni, H. Motalebi, B. Minaei-bidgoli, M. Shokrollahi-far, A farsi part-of-speech tagger based on markov, *In the proceedings of ACM symposium on Applied computing*, Brazil (2008).

[7] S. Jabbari, B. Allison, Persian Part of Speech Tagging, *In the Proceedings of Workshop on Computational Approaches to Arabic Script-Based Languages (CAASL-2)*, USA (2007).

[8] E. Brill, Transformation-Based Error-Driven Learning and Natural Language Processing: *A case Study in Part of Speech Tagging, Computational Linguistics*, USA (1995).

[9] M. Hepple, Independence and Commitment: Assumptions for Rapid Training and Execution of Rule-based Partof-Speech Taggers, *In Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL)*. Hong Kong (2000).

[10] K. R. Singha, B. S. Purkayastha, K. D. Singha, Part of Speech Tagging in Manipuri with Hidden Markov Model, *IJCSI* International *Journal of Computer Science Issues*, Vol. 9, Issue 6, No 2, November 2012

[11] M. Okhovvat , B.M. Bidgoli. A Hidden Markov Model for Persian Part-of-Speech Tagging, *Procedia Computer Science* 3 (2011) 977–981

[12] A.J.P.M.P. Jayaweera, N.G.J. Dias. Hidden Markov Model Based Part of Speech Tagger for Sinhala Language, *International Journal on Natural Language Computing (IJNLC)* Vol. 3, No.3, June 2014

[13] D.L.Cing, K.M.Soe. Joint Word Segmentation and Part-of-Speech (POS) Tagging for Myanmar Language, *17th International Conference on Computer Application*, Yangon, 27-28, February,2019

[14] https://github.com/ye-kyaw-thu/sylbreak

BN 978-981-14-4787-7
Proceedings of 2020 the 10th International Workshop on Computer Science and Engineering (WCSE 2020)
Yangon (Rangoon), Myanmar (Burma), February 26- February 28, 2020, pp. 68-72

A Fear State Judgement System for Alleviating Fear of Heights Gradually in VR

Iku Kitanosono[1+], Toshiyuki Haramaki[2] and Hiroaki Nishino[3]

[1]Graduate School of Engineering, Oita University, Japan
[2, 3]Faculty of Science and Technology, Oita University, Japan

Abstract. VR has been expanding into various fields in recent years. The medical field is no exception to this trend. For example, VR technology is expected to be used to treat acrophobia, the most common phobia. There are many patients with acrophobia, and many of them have difficulties in their daily lives. One way to solve this problem is to treat acrophobia with VR. However, VR treatment is a big problem in the treatment of phobia because there are many individual differences. In this study, we propose a system that can acquire the fear state of an individual in real-time by vital sensing and can treat phobia of altitude gradually adjusted to the individual using a placebo effect.

Keywords: acrophobia, VR, medical treatment, vital sensing, placebo effect

1. Introduction

In recent years, VR has made inroads into various fields, and research is underway in various places. The medical field is no exception to this trend and increasingly used to treat phobias. Although it is normal for people to feel a sense of fear about their own life and death, patients with acrophobia may have negative feelings about heights, which can damage their daily lives and the people around them. Therefore, it was considered desirable to develop simulation software that can easily treat acrophobia using VR. VR systems for the treatment of acrophobia already exist, for example, in games and researches. However, all such software products are highly dependent on individual differences. From this viewpoint, we constructed a system that can provide more effective treatment for fear of heights after an approach to fill in individual differences using vital sensing and placebo effects. The development was carried out by dividing it into the VR system which treats fear of heights using the placebo effect and the system which can judge the fear condition of the users.

2. Related Works

We have developed a VR system to treat acrophobia using the placebo effect [1]. The objective is to treat light acrophobia. A user who wants to treat acrophobia writes a consent form to immerse into this system. Then, a user equips a VR device and board an elevator to the room which is located in high place. First, the user experiences the second floor and go upper gradually. We established the transparent level as terms to clear each floor. Transparent level means a method to make users understood the room locates high places gradually by experiencing level 1 to 3 of transparent level. Level 1 transparent window frame and cell, level 2 transparent wall, and level 3 transparent floor. This is also a placebo effect. This system defined that placebo effect means the thing which users get used to high places gradually, so user clear each transparent level and misunderstand that he is in high places visually.

[+] Corresponding author. Tel.: + 81 070-2667-6516
E-mail address: v18e3009@oita-u.ac.jp

From the result of the experience, it is effective to get used to high places gradually, but it is weak the effect to use the placebo effect. Therefore, we propose a more effective method of improving phobia by combining the fear state judgment system described in this paper with the placebo effect.

3. Methodology

3.1. Summary

As described in 2, a VR system for treat fear of heights needs to determine the fear state. So, to indicate the fear state visually, this study determined the fear state by mapping the acquired vital data of the user onto Russell's emotional model [2]. Various famous models have been proposed so far. However, since Ekman's basic facial expression [3] is for facial expressions, it is inappropriate for this system where most of the face is hidden in VR. Besides, the method of mapping the vital sensed data seems to be difficult in the circle of the feeling model put forward by Plutchik. Therefore, we used Russell's ring model, which consists of 2 opposing axes. Fig.1 shows Russell's emotional model, that contains of the x-axis for comfort and discomfort and the y-axis for arousal and sleepiness. The state of fear is determined when the patient is in an arousal and discomfort state located in the second quadrant.

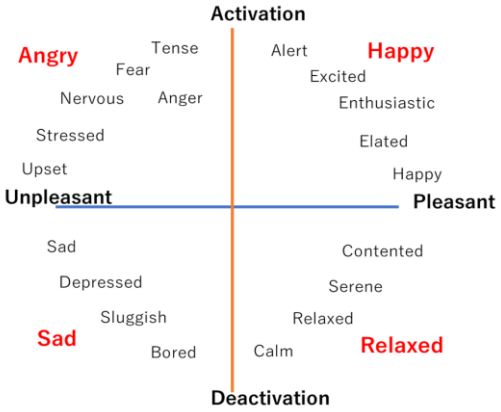


Fig. 1: Russell's emotional model

3.2. Mapping on Russell's Emotional Model

To mapping on Russell's emotional model, the Pleasant/Unpleasant axis is acquired by EEG, and the Activation/Deactivation axis was acquired by ECG. It is possible to get relaxation or concentration data from EEG. In addition, it is known that drowsiness also appears in ECG.

The fear or tension state is in the second quadrant on Russell's emotional model and it is necessary to detect an Unpleasant state on the x-axis and Activation state on the y-axis. Since the Unpleasant state is acquired from the EEG, the degree of relaxation can locate as the x-axis component. Moreover, regarding the Activation state, it is necessary to measure the activity level of the sympathetic nerve. The sympathetic nerve is a nerve in a tension mode that is active during the day, and the parasympathetic nerve is a nerve in a relaxation mode that is predominantly active at night. For this reason, the sympathetic nerve becomes active in the awake state, and the parasympathetic nerve becomes active in the sleepy state.

The ECG has a period and the part with the largest amplitude within the period is called R. The interval between R and R is called the R-R interval (RRI).

$$\bullet \quad rMSSD = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N-1}(x_i - x_{i+1})^2} \tag{1}$$

rMSSD is a parameter representing the mean square root of the square of the difference of RRI and representing the activity state of the parasympathetic nerve.

$$\bullet \quad SDNN = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{2}$$

SDNN is the standard deviation of RRI. This is a parameter that represents the sympathetic and parasympathetic activity states. Since this ratio can be used to determine the state of sympathetic activity, it is necessary to check whether the SDNN / rMSSD value has increased.

4. System Implementation

4.1. Mapping to the Axis Representing Pleasant/Unpleasant

EEG is used to determine the position along the pleasant and unpleasant axis. Our system used a MindWave Mobile 2 [4] which is developed by NeuroSky Co. to acquire EEG. EEG data is sent and received via Bluetooth connection with PC. Since there was only a Python 2.7 library that could receive data from the device, we created a Python 2 program that receives the data, sends it over a socket to localhost, and then creates a Python 3.6 program that receives the data and summarizes it with heart rate sampling data on the y-axis. The data sent from the device contains various values such as α waves and round meditation, but this time we used meditation value.

4.2. Mapping to the Axis Representing Activation/Deactivation

The ECG was utilized to determine the position along the activation and deactivation axis. The used device was an Arduino Uno DFRobot Heart Rate Monitor Sensor [5] with electrodes. Using Arduino language developed based on C language, we created a program that continuously sends RRI through this device. As a method of detecting the R part, the maximum value was continuously updated, and the peak value was determined when it continuously dropped to 70 percent of the maximum value, and the same was done for each period to obtain the difference. Peak value means R value. And the maximum value means the maximum part for the interval. We also created a Python 3.6 program that receives RRI data and samples data as SDNN / rMSSD sent via serial communication.

4.3. Draw to Graph on Russell's Emotional Model

Fig. 2 shows the scatter graph which maps each acquired data. A real-time continuous data obtained in and **4.1** and **4.2** are mapped onto the graph which indicates Russell's emotional model. In the Fig. 2, the vertical and horizontal scales are different from Russell's emotional model, but they are the same. As the x-axis means **4.1** and the y-axis means **4.2**. The mapping is done using the matplotlib Python library. Currently, UI is implemented using the scatter chart, but we plan to improve it to make it easier for users to understand.

The origin is (x-axis = 50, y-axis = (SDNN / rMSSD)). The x-axis origin indicates median relaxation 50 degree which has 0 – 100. The y-axis origin is drawn with SDNN / rMSSD value excluding the user's outlier immediately after the system is started as the reference value. When the x-axis component becomes a value of 50 or less, it indicates an Unpleasant state, so that it is located on the quadrant of x<=50 in the graph. Further, when the value of the y-axis component is equal to or less than 50, it indicates an uncomfortable state, and therefore, the graph is positioned in the quadrant of x < = 0. When it is satisfied with the above two conditions, a value located in the second quadrant is obtained.

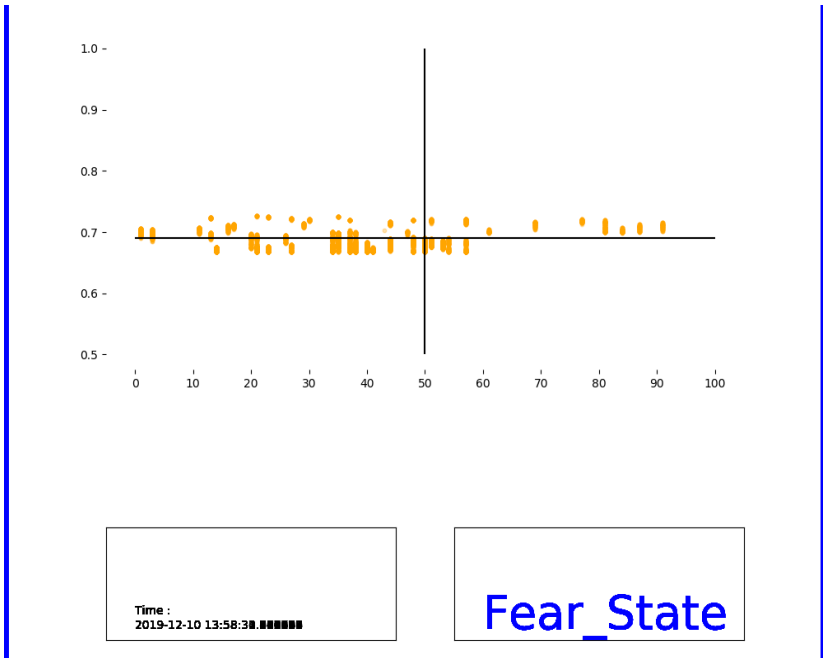


Fig. 2: This system's UI for drawing Scatter Graph which indicates Russell's Emotional Model

5. Preliminary Experiment

5.1. Outline of the Experiment

We experimented to watch a movie which felt fear, because this system is made for judging fear state. It is necessary to acquire a fear state that the individual does not perceive, so it is necessary to perform a quantitative evaluation instead of a qualitative evaluation that depends on the subject's subjectiveness. Therefore, it is necessary to use an emotion measurement device different from this system for quantitative evaluation. We used the system called “Affedex” which was an AI to recognized human emotions. Affedex is developed by CAC company. In this case, we used the free Android app “Affedex me”, because we needed to conduct experiments to evaluate the accuracy. We would like to consider the results of experiments conducted using the system for overcoming acrophobia, but since Affedex me is a system that infers emotions from the user's facial expressions, VR that covers the entire face not compatible with. Therefore, we asked the participants to experience the video. After this experiment, we wanted to have high-altitude phobia patients experience high-altitude videos and evaluate them.

5.2. The Environment of the Experiment

Three students wearing Mind Wave Mobile 2 and Arduino Uno DFRobot Heart Rate Monitor Sensor watched a 9-minutes horror video which includes elements that surprise you. The fear state of the student while watching the fear video is saved with the time data and all the ECG data and EEG data while watching the fear image are saved in a CSV file.

Emotions are estimated using Affedex me from the expression of the subject while watching a video. In order to perform quantitative evaluation, the results of this system and Affedex me are compared and evaluated.

5.3. Result of the Experiment

Fig. 3-5 shows the result of fear state detected from User A-C. The arrow at the side of “Horror Video” shows the time of the video. This system shows the fear state of User A to C which detects by this system. Affedex me shows the fear state of User A to C which detects by Affedex me system.

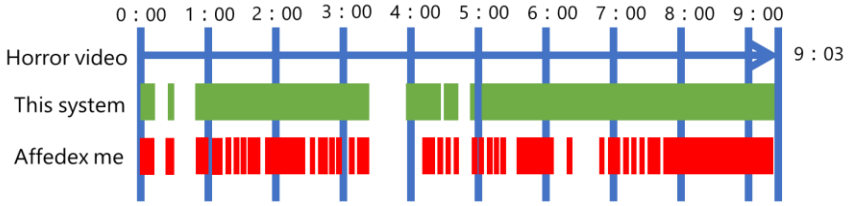


Fig. 3: User A's fear state compared to Affedex me

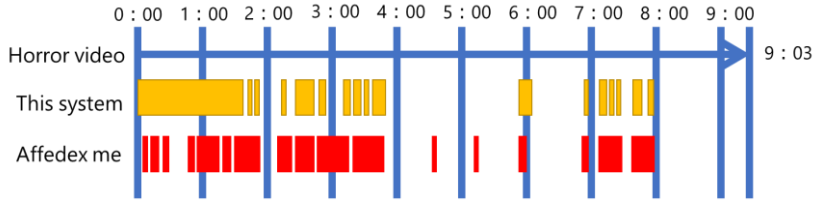


Fig. 4: User B's fear state compared to Affedex me

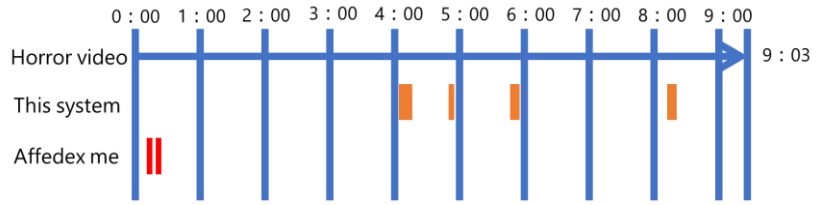


Fig. 5: User C's fear state compared to Affedex me

5.4. Consideration

From the results of Fig. 3 and 4, it was confirmed that the fear states of the users A and B were normally acquired. However, as a result, Affedex me can detect the state of fear in more detail, so it is necessary to consider it as a future improvement. From the results of Fig. 5, a fear state that did not match at all was detected. Judging from the answer that the user C did not feel any fear when he/she checked with the user C after the experiment, it is considered that he/she erroneously recognized the fluctuation of vital data such as the movement by the device as the detection. In the future, I would like to be able to acquire data stably even when it operates by re-examining devices.

6. Conclusion

This paper proposed a fear state acquisition system for more effective treatment of acrophobia with VR. Then, we devised and implemented the algorithm, and conducted experiments to confirm and improve the accuracy of the fear state of this system. In the future, this system will be able to predict the state of fear and judge with high accuracy by introducing RNN using deep learning. Prediction reduces the mental and physical burden and allows for more gradual treatment. We also think that if we can predict emotions, we can introduce not only acrophobia but also various system.

7. Acknowledgements

We would like to express my appreciation to Mr. Gyohten who assisted the submission to this conference.

8. References

[1] Iku Kitanosono, Toshiyuki Haramaki, Tsuneo Kagawa, Hiroaki Nishino: A VR System for Alleviating a Fear of Heights Based on Vital Sensing and Placebo Effect. International Conference on Network-Based Information Systems (NBiS) 2019: Advances in Networked-based Information Systems pp 62-72

[2] James A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology.* Vol. 36, pp. 1161 ~ 1178. 1980.

[3] P. Ekman and W.V. Friesen : Facial Action Coding System, Consulting Psychologist Press. 1977.

[4] NeuroSky MindWave Mobile2(https://www.neurosky.jp/mindwave-mobile2/)

[5] DFRobot Heart Rate Monitor Sensor(https://wiki.dfrobot.com/Heart_Rate_Monitor_Sensor_SKU__SEN0213)

Chapter 3
Image Processing

ISBN 978-981-14-4787-7
Proceedings of 2020 the 10th International Workshop on Computer Science and Engineering (WCSE 2020)
Yangon (Rangoon), Myanmar (Burma), February 26- February 28, 2020, pp. 73-78

Persistent Operation of OF@TEIN+ Playground Verified by SmartX Multi-View Leveraged Visualization

Muhammad Ahmad Rathore[1], SeungHyung Lee[1] and JongWon Kim[1 +]

[1] School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju 61005, Korea

Abstract. In OF@TEIN+: Open/Federated Playgrounds for Future Networks (SDN/NFV/Cloud-integrated) testbed, the persistence operations of infrastructure is extremely critical to ensure availability of physical/virtual resources and identify key performance bottlenecks, which can degrade testbed operations. Persistence-enabled operations are increasingly evident in testbeds, and as a result, there is a growing focus on monitoring its long-terms effectiveness and effectively utilizing the deployed services, enabling end-users to reliably utilize playground applications. Significant challenges in measuring persistence, however, contribute to both a tendency towards identifying efficient metrics with resilient functionalities and effective verification methods. In this paper, we present our idea of maintaining persistent operation of OF@TEIN+ playgrounds and verify through visualization of gathered operational data.

Keywords: Persistent operations, multi-layer visibility, visualization, cloud-native

1. Introduction

Maintaining a steady state of monitoring has always been a key element in ensuring the sustainable performance of complex distributed systems, being the first step to control quality of service, detect failures, or make decisions about resource allocation. Furthermore, enabling end-users (developers, etc.) for effectively utilizing the available services is crucial for persistent operations. Aligned with Future Internet testbeds, we launched OF@TEIN+ in 2017, to further extend and expand OF@TEIN collaboration [1]. Fig 1 shows, OF@TEIN+ multisite cloud (denoted as OF@TEIN+ Playground) connects around 10 international sites in 10 countries (Korea, Malaysia, Thailand, Indonesia, Laos, Cambodia, Vietnam, Myanmar, Bhutan, and India) and interconnected via OF@TEIN+ network. In the OF@TEIN+ project, multi-site affordable Playground sites are established which consists of hyper-converged box-style resources named (Type O) "SmartX Micro-Box" having interfaces for management/control and data and IoT devices. SmartX Micro-Box is configured to support Cloud-native computing with containerized IoT-SDN-Cloud functionalities and capabilities of edge computing that allow computation to be performed at the network edge near data sources [2].

OF@TEIN+ Playground as shown in Fig. 1 supports multiple resource types: physical, virtual, and container types [3]. In the multisite cloud, infrastructure burdens of large volume generated by end-users or devices and transferred to a centralized cloud lead to inefficient utilization of bandwidth, storage and computing resources [4]. In order to effectively operate OF@TEIN+ Playground, it is truly requisite to recognize how resources (physical/virtual/container) are running over a time period. Furthermore, maintaining the knowledge of networking connection, through monitoring and visualization is essential to understand and troubleshoot server and network issues before they affect the end-users. Lately, we have been developing tools for monitoring and measurement of OF@TEIN+ Playground resources enabled with cloud-native edge computing that delivers a rich understanding of operations through visibility data and provides

+ Corresponding author. Tel.: +82-62-715-2219
E-mail address: jongwon@gist.ac.kr

deep insights into current operations, without worrying about underlying infrastructure. By applying persistent visibility to OF@TEIN+ Playground, leveraging modified 'SmartX Multi-View Visibility Framework (MVF)' we can monitor various dynamic visibility metrics from multiple measurement points across both physical and containerized resources and associated flows of the playground [5]. However, "persistent operations" deal effectively with managing playground resources leveraging visibility data as well as enabling end-users to utilize playground applications in a stable and reliable way. In addition, visualization of playground operations enables the operators to make informed decisions to tackle the above-mentioned challenges.

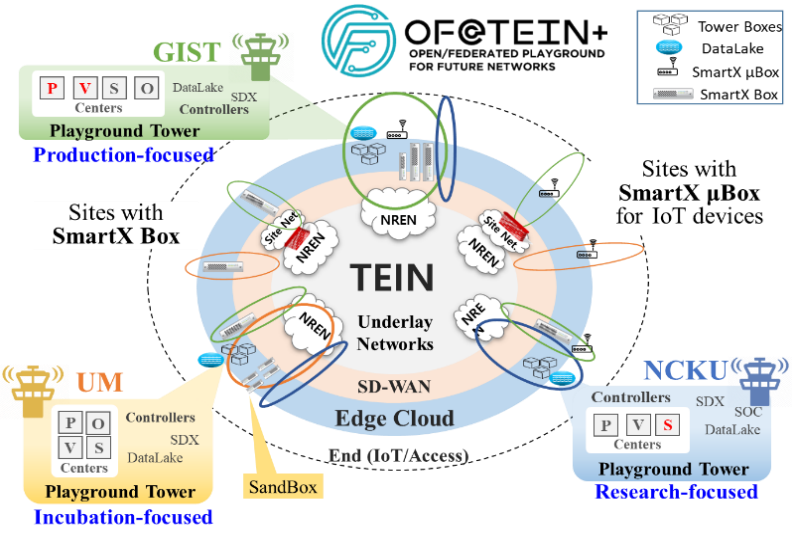


Fig. 1: OF@TEIN+ Playground as multisite clouds.

Thus in this research, leveraging multi-layer visibility data and supporting efficient usage of playground applications, we propose a solution for persistent operations of playground with visualization support for verification. In summary, the key contributions of this paper are:

- We formulate the requirements for maintaining visibility data together with utilizing the cloud-native applications.
- Present design and implementation for maintaining resilient and robust collection of visibility data, followed by integration (i.e. prepared and inferred) at a centralized place.
- Present design and implementation for employing cloud-native containerized applications, distributed across playground.
- Verify persistent operations of playground with visualization schemes such as multi-belt onion-ring visualization.

The remainder of this paper is organized as follows. In Section II, we briefly discuss requirements. In Section III, we collectively cover both design and preliminary prototype implementation of achieving persistent operations. After that, we discuss verification results in Section IV. Finally, section V concludes the paper.

2. Background and Related Work

2.1. SmartX Multi-View Visibility for Persistent Operation

In OF@TEIN+ playground persistent monitoring/ measurement leveraging SmartX MVF is proposed to deal with the multiple layers of visibility and visualization such as resource layer (underlay, physical), flow-layer, and workload-layer [6]. In resource-layer visibility, monitoring and visualization of playground physical resources and inter-connects (e.g. paths and links) are considered. Whereas flow-layer visibility monitors overlay network traffic in near real-time through packet tracing. A network flow is typically a sequence of network packets that belongs to a certain network sessions between two endpoints. Flow layer visibility deals with different levels of flow information (i.e., collected, clustered, identified and un-clustered

flow) by utilizing a balanced flow collection, clustering, and tagging. Next, workload-layer visibility is responsible for monitoring and visualization of inter-connected containerized functions (e.g. Web, App, DB).

2.2. Limitations of Existing Tools

A number of open-source visibility monitoring and visualizations tools already exists such as Vendetta [6] for monitoring distributed testbed, monitoring of OpenFlow-based SDN [7] and Monarch based on Software-defined infrastructure [8]. However, these tools do not directly cater to persistent operation and intuitive visualization for single-box-virtualized deployment style of OF@TEIN+ playground. In order to solve this limitation, we propose a unified persistent monitoring solution to monitor entire infrastructure including heterogeneous resources (e.g. physical, virtual, and container), inter-connects and applications (infrastructure). The initial design of SmarX MVF applies network-packet precise collection at Micro-Box, and then this raw collection is transferred at the Visibility center to perform the processing for flows creation. This process induces burden on storage capacity and network bandwidth. To handle these issues, we introduced the process of data aggregation, tagging to crate flows at the Micro-box instead of doing the same at visibility center.

2.3. Requirements of Persistence Operations for OF@TEIN+ Playground

Maintaining persistent visibility in cloud-native infrastructure poses several challenges such as Micro-Boxes may join and leave the network (e.g. connection failure). Multi-view visibility is required to understand downtime, and manage physical assets. Secondly, job scheduling, resource provisioning, and allocation mechanism require monitoring metrics with near real-time collection and low latency. However, in a graphically distributed infrastructure edge Boxes could induce delay. This creates a need for keeping the monitoring data at the edge and keeping the resource collection at a reasonable volume. Thirdly, in edge devices, for reliable transfer of visibility data at the Visibility Center, a local/temporary storage is required during disconnections, unexpected applications closure. Lastly, for faster troubleshooting, visualization must flexibly incorporate multi-layer visibility data, collected from multiple sites with multi-tenancy support.

3. Persistence Operations of OF@TEIN+ Playground: Design and Porotype Implementation

In this section, we present design and preliminary prototype implementation of proposed persistent operation.

3.1. Selected Tools for Persistent Operation

The persistent operations of visibility collection require the identification of key monitoring tools that sustains major impact on operational states of the playground . For resource layer visibility collection we utilize *Collectd*[1] to collect performance statistics of Micro-Box. For flow-layer visibility collection, we use *eBPF*-based packet tracing tools. While *Apache-Spark*[2] with Scala is utilized to generate flows from packet tracing. Besides, for resource-layer visibility we employed *PerfSONAR*[3] command-line tools for monitoring playground resources, interconnects and end-to-end network performance metrics. To ensure consistent running of monitoring applications we placed Box-agents at each resource that periodically check the running status and actively start the application within a short interval. Besides, Box-Agents communicate with Centre-Agent through *zeroMQ*-based communication to handle asynchronous communication. At the visibility Center together with Apache Kafka[4], to manage reliable and persistent transport of visibility data, Apache *ZooKeeper* provides automatic management of metadata and synchronization issues. A customized java-based tool parse and validate the visibility data before storing it in the appropriate databases (i.e., *MongoDB*[5], *ElasticsSearch*[6] and *InfluxDB*[7]). A scheduler aggregates values of measurements at end of the

[1] collectd collects metrics from a number of sources, e.g. the operating system, applications and stores this information over the network
[2] Apache Spark is an open-source fast and general engine for large-scale data processing.
[3] perfSONAR is a test and measurement infrastructure that is used by science networks around the world to monitor and ensure network performance.
[4] Apache Kafka is an open-source distributed stream processing platform for building real-time data pipelines and streaming apps written in Java etc.
[5] MongoDB is an open-source document database that provides high performance, high availability, and automatic scaling.
[6] Elasticsearch is an open-source, multi-tenant, distributed, search engine for cloud environment with various set of APIs.
[7] InfluxDB is an open-source time series database useful for recording metrics, and performing analytics with built-in HTTP API support.

day to be useful for analysis purposes. For single view unified visualization, we apply multi-belt onion-ring visualization together with Kibana and Grafana. While, for orchestration of containerized application, we employ Kubernetes.

3.2. Design and Porotype Implementation

We leverage distributed multi-site SmartX Micro-Box supporting cloud-native (containerized) IoT-Gateway and prepared as Kubernetes-orchestrated workers with SDN-coordinated special connectivity. At the Micro-Box the p*ersistent visibility collection and validation* stage collects and validates monitoring and measurement data based on selected visibility metrics. Formatted visibility data is sent to *Persistent Visibility Storage* and *Staging* stage where data is stored. Next *Persistent Visibility Integration* stage integrates collected data for generating consolidated reports over a period and identifies any anomalies. Finally, *persistent Visualization* stage accesses processed data and auto-generates graphical views. As shown in Fig 2, we deploy SmartX-Micro-Box at multiple Sites in OF@TEIN+ Playground with capabilities of sending control/data messages at the Visibility Center levering multiple tools. For utilization of playground operations by end-user, we developed a cloud-native service through Kubernetes and distributed it through OF@TEIN+ playground. One Kubernetes cluster was formed with Edge IoT-gateways distributed at Multi-site of OF@TEIN+ playground.

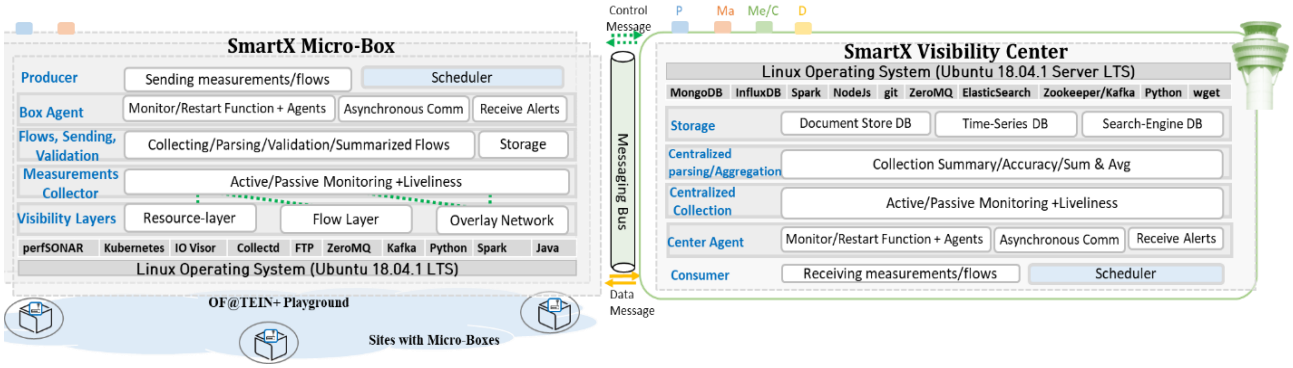


Fig. 2: Design of persistent visualization for SmartX Multi-View Visibility.

4. Persistent Operation of OF@TEIN+ Playground: Verification

4.1. Persistent Visibility Operation

For verifying the prototype implementation, we utilize the OF@TEIN+ playground. We consider metrics from visibility tools and visualized visibility data at each step of visibility workflow.

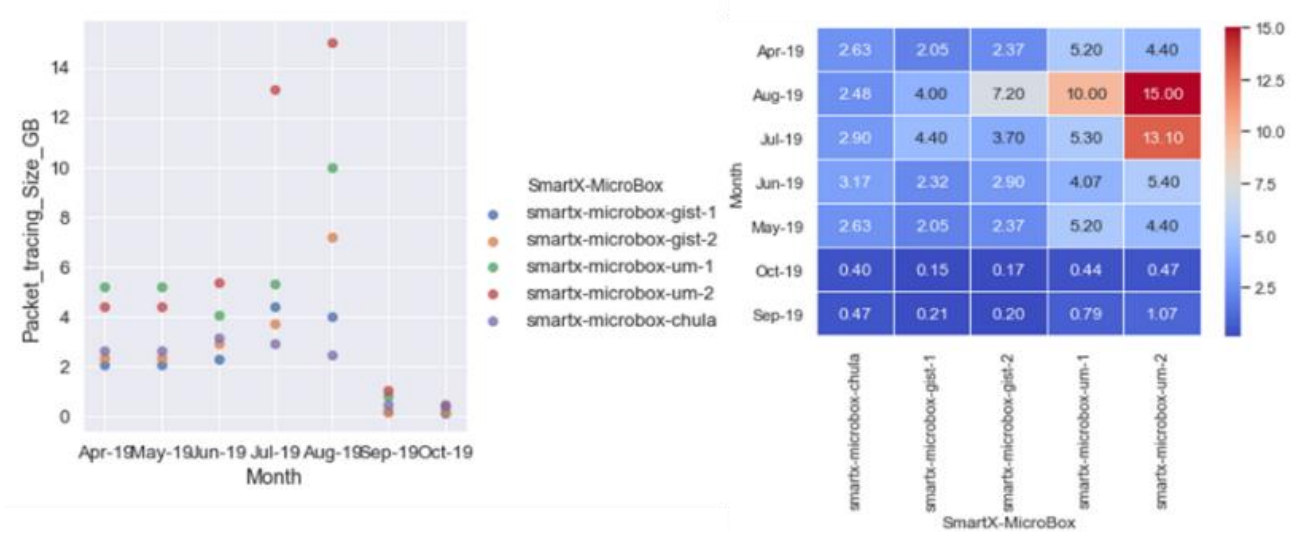


Fig. 3: Visualization for generated volume during packet tracing

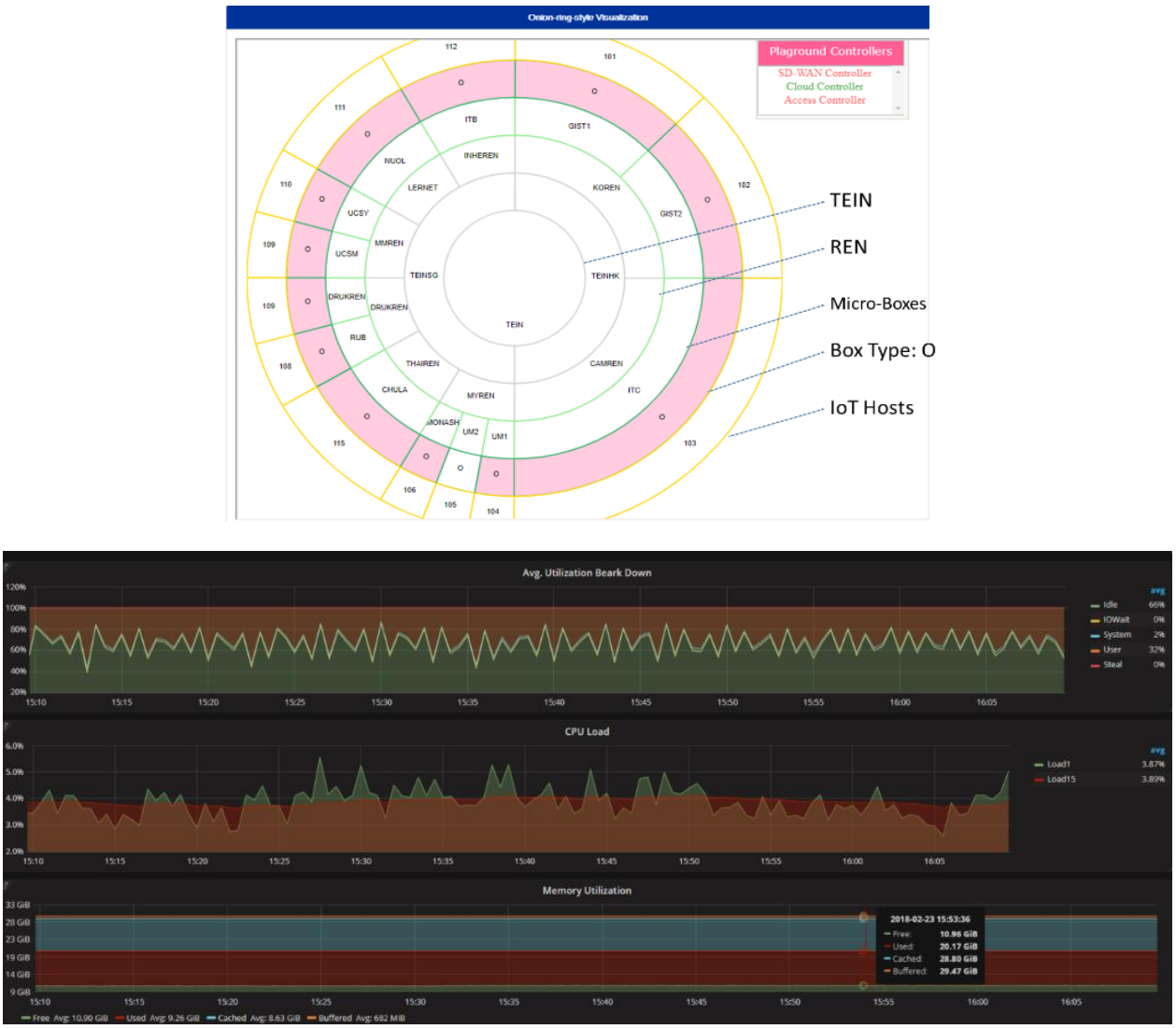


Fig. 4: Onion-ring visualization of OF@TEIN+ Playground (top), Grafana-based visualization of Resource Layer (down)

Figure 3 shows verification for reduction in collection size from previous months. In the last two months (i.e. Sept-Oct 19) where flows are generated at the Micro-box, size of visibility data was reduced by over 80% comparing with previous months (i.e. Apr-Aug 19). Fig. 4 (top) shows a multi-belt onion-ring visualization, which is an effective dashboard to present underlay resource-layer, physical resource-layer, and flow-layer visibilities together in a single unified view. Onion-ring supports interactive visualization, i.e. clicking on layer labeled “Micro-Box”, opens a sub-window that shows visualization for Physical resource layer with multiple metrics leveraging Grafana and views its relationship with the underlay networks as shown in Fig. 4 (down).

4.2. Cloud-native Service Operation

For verification of operations usage, we develop applications on a cloud-native platform i.e. Kubernetes to facilitate in terms of automatic deployment, scaling, and management of distributed cloud-native applications. These applications support a service that combines IoT and cloud using SmartX framework. Thus, cloud-native service operation is verified through Smart energy service, as SmartX IoT-Cloud Service. Persistent operation needs to understand the current state of the service functions and make sure they are working well in isolation without conflicting with other containers. In this paper, we check the status of service containers deployed on OF@TEIN+ Playground using Weave scope, an open source visualization tool, to verify that services are normally deployed and running. As shown in Figure 5, we verified that smart energy IoT-Cloud service is well running by using the cloud-native Kubernetes cluster deployed in multi-site OF@TEIN+ Playground by monitoring the operation status of the containers that make up the service.

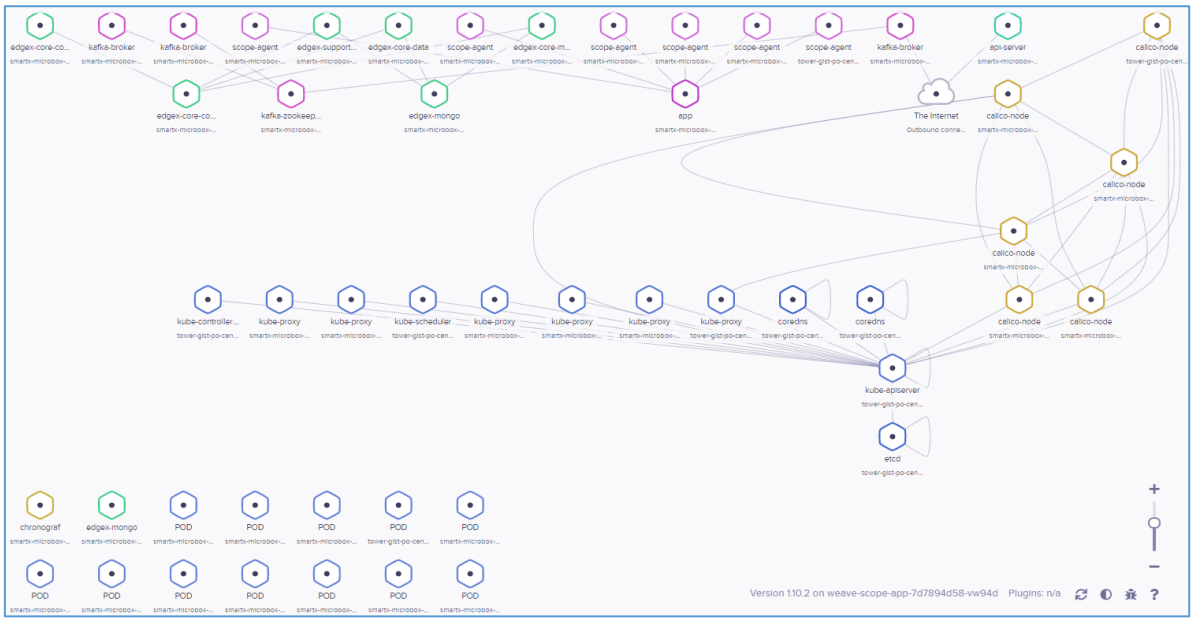


Fig. 5: Smart energy service containers deployment in OF-TEIN+ playground

5. Conclusion

In this paper, we presented our initial effort to provide persistent operations for OF@TEIN+ playground developers and operators to effectively operate and maintain the playground. We verified the work by persistently visualizing multiple layers of visibilities leveraging SmartX Multi-view visibility.

6. Acknowledgments

This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2017-0-00421, Cyber Security Defense Cycle Mechanism for New Security Threats). This work is also partially supported by the Data-centric IoT-cloud service platform for smart communities (IoTcloudServe@TEIN) project under the WP4 Future Internet of Asi@Connect.

7. References

[1] Kim J, Cha B, Kim J, et al. OF@ TEIN: An OpenFlow-enabled SDN testbed over international SmartX Rack sites. *Proceedings of the Asia-Pacific Advanced Network* 2013; 36: 17–22.

[2] Shi, Weisong, and Schahram Dustdar. "The promise of edge computing." Computer 49.5 (2016): 78-81.

[3] Usman, Muhammad, Nguyen Tien Manh, and JongWon Kim. "Multi-belt Onion-ring Visualization of OF@ TEIN Testbed for SmartX Multi-View Visibility."

[4] Usman, Muhammad, et al. "SmartX multiview visibility framework leveraging open-source software for SDN-cloud playground." 2017 IEEE Conference on Network Softwarization (NetSoft). IEEE, 2017.

[5] M Ahmad Rathore, M Usman, JW Kim. “Maintaining SmartX Multi-View Visibility for OF@TEIN+ Distributed Cloud-native Edge Boxes.” Transactions on Emerging Telecommunications Technologies (2019): Submitted.

[6] Rensfelt, Olof, Lars-Ake Larzon, and Sven Westergren. "Vendetta-a tool for flexible monitoring and management of distributed testbeds." 2007 3rd International Conference on Testbeds and Research Infrastructure for the Development of Networks and Communities. IEEE, 2007.

[7] Isolani, Pedro Heleno, et al. "Interactive monitoring, visualization, and configuration of OpenFlow-based SDN." 2015 IFIP/IEEE International Symposium on Integrated Network Management (IM). IEEE, 2015.

[8] Lin, Jieyu, et al. "Monitoring and measurement in software-defined infrastructure." 2015 IFIP/IEEE International Symposium on Integrated Network Management (IM). IEEE, 2015.

ISBN 978-981-14-4787-7
Proceedings of 2020 the 10th International Workshop on Computer Science and Engineering (WCSE 2020)
Yangon (Rangoon), Myanmar (Burma), February 26- February 28, 2020, pp. 79-83

Effective Multi-View for Human Activity Recognition on Skeletal Model

Sandar Win [1], Thin Lai Lai Thein[2+]

[1 2]University of Computer Studies, Yangon, Myanmar

Abstract. The recognition of 3D human pose from 2D joint location is fundamental to numerous vision issues in analysis of video sequences. Various methods using with skeletal model have been described in past decades, but there is required a powerful system with stable and reliable manner in activity recognition because video sequences can contain different people that may be any position or scale and complex spatial interference. With the development of deep learning, skeleton-based human representation is more reliable to motion speed and appearance of human body scale. Skeleton data contains compact information of the major body joints and that support multi-view to human activity recognition. To satisfy our aim, the proposed system is developed by using OpenPose detector that achieve effective results for 2D pose and Deep Learning based approach. Our goal is to extract valuable information between human joints and to recognize correct activity from human representation in video sequences.

Keywords: OpenPose, Human Activity Recognition, Deep Learning

1. Introduction

The recognition of perception visual data in human activity is generally analysed into five categories that are body part locations or skeletal joints, 3D silhouettes, local space-time information, features of scene flow, and local residence features [1]. Skeleton-based human activity recognition has affected in a great deal of interest to a lot of researchers and available in real-world applications such as intelligent surveillance, activity recognition, sport video analytics, and autonomous monitoring system, etc. Currently, several approaches were developed by using 3D skeletal joint positions that is directly taken from sensors. Many of researchers used depth images to recognize human activity in their system. Depth image provides geometric feature of pixel information that supports the scene in 3D space and estimate 3D pose based on depth information, but computing of depth maps is slow and prone to errors when searching the correspondence map in noisy depth information [2]. On the other hand, by using depth images and reconstructing of 3D point clouds are robust to scale, rotation and illumination changes [3]. This approach is very constraint in application area. Another approach is automatic acquisition of accurate 3D pose from an image requires a very sophisticated setup. And, some approach calculates relative joint orientations and utilizes order of joint to connect adjacent vectors [4], but occurs ambiguity in recognition system. So, effective multi-view for activity recognition from video sequence have been required. To solve these problems, the system integrates OpenPose with Joint Correlation Distance and skeleton visualization method to estimate human pose for activity recognition. The purpose of our system is to propose skeleton based robust method on the Deep Learning Unified framework. The remained portions of the paper are organized as follows: Related works are described in Section 2. Methodology of proposed approach is presented in Section 3. Experimental results are shown in Section 4, and conclusion and future works are expressed in Section 5.

2. Related Works

[+] Corresponding author. Tel.: + 95-09-793757403; fax: + 95-013-610-633.
E-mail address: sandarwin@ucsy.ed.mm

The skeleton data have been widely utilized for activity recognition system because it can provide dynamic conditions and complex circumstances. Since human joint information in skeleton data have been proved great success for action recognition tasks. Wu and Shao [5] proposed to recognize human action by using deep neural networks with hierarchical dynamic framework on extracted 3D skeleton feature and estimate the probability of action sequences. Luvizon et al. [6] proposed a multitask framework for 2D joint and 3D pose estimation from video sequences in recognition of human action. They trained multi type of dataset to generate 3D predictions from 2D annotated data and proved an efficient way for action recognition based on skeleton information. Iqbal et al. [7] developed dual source network on 3D human pose estimation. They collected large amount of unconstrained data in 2D and 3D pose. And by taking nearest 3D pose and reconstructed the 3D pose for single image estimation. Du et al. [8] proposed Recurrent Pose Attention Network (RPAN) that predicts the related features in human pose. The system used end to end recurrent network layers for temporal action modelling to construct a skeleton-based human representation. As the number of layer increase, the representations extracted by the subnets are hierarchically fused to figure a high-level feature to represent human in 3D space. Yang et al. [9] proposed Double-feature Double-motion Network (DD-Net) to achieve lightweight network structure by employing skeleton sequence attributes and motion scale variation. Li et al. [10] described translation scale invariant method that work well on 2D skeleton video. They used Convolutional Neural Network (CNN) architecture and result is compared on benchmark dataset. In activity recognition system, most of system used depth image. Although these are very constraint in outdoor applications. To eliminate this constraint, the proposed system is developed with efficient multi-view for human activity recognition from video sequences which contains forward, frontal, lateral and backward motion.

3. Methodology

The proposed system consists of three parts. There are body pose detection from video sequences and extract 2D joint locations, 3D pose estimation and Activity Recognition on Deep Neural Network. An overview of proposed system is shown in Fig.1.

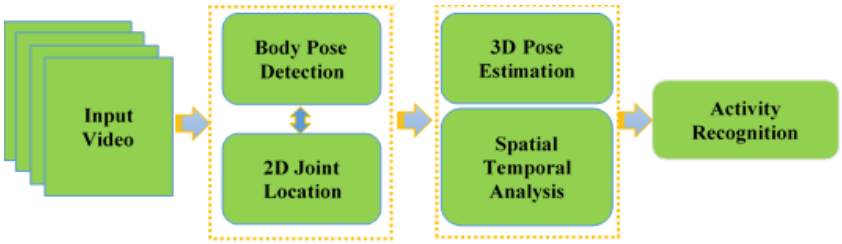


Fig. 1: An overview of proposed system. Input video is first mapped with OpenPose detector via activity recognition on Deep Neural Network

3.1. Body Pose Detection and Extract 2D Joint Locations

The system detects human body by OpenPose detector that search key points of human body parts and extract 2D joint locations. To perform better result, a non-maximum suppression method is used to select the interest points in the representation structure and that returns the top scoring result as defined the estimated pose. This approach also supports partial occlusions and improves motion detection [11].

3.1.1. Scale Invariant Motion Modeling

For skeleton-based action recognition, geometric feature and Cartesian coordinate feature are used as an input feature. The geometric feature such as joint indices and distances are location viewpoint invariant, but not stable from one data to another. The indices of joint (i.e., the IDs of the head, left and right shoulders, etc.) could be dynamically changed in different actions. Hence, the difficulty arises and requires to predefine the correlation of joints by ordering of their indices. By using Joint Collection Distances (JCD) can solve for problems. As another benefit, the embedding process can reduce the effect of skeleton noise. Joint collection as defined by $S_k = \{J_1^k, J_2^k, \ldots, J_N^k\}$, where k is number of frame and JCD feature of $S_k$is computed:

$$JCD^k = \begin{bmatrix} \left\|\overrightarrow{J_2^k J_1^k}\right\| & \cdots & \left\|\overrightarrow{J_2^k J_{N-1}^k}\right\| \\ \vdots & \ddots & \vdots \\ \left\|\overrightarrow{J_N^k J_1^k}\right\| & \cdots & \left\|\overrightarrow{J_N^k J_{N-1}^k}\right\| \end{bmatrix} \quad (1)$$

where $\left\|J_i^k J_j^k\right\| (i \neq j)$ denotes Euclidean distance between $J_i^k and J_j^k$

3.2. 3D Pose Estimation

Pose estimation is an important class in action recognition. 3D pose estimation from 2D joint locations with skeleton data has advantage to view point invariant and greatly effects on performance. We also consider changes between 2D and 3D pose structure to reduce ambiguity physical constraint on limb lengths. Deep fully convolutional network is trained to predict the uncertainty maps of the 2D joint locations. Then, 3D pose is estimates via an Expectation-Maximization (EM) algorithm over the entire sequence. EM algorithm search probability distribution of 2D joint location (heat map value) in frame t and compute mean and normalized nearest 3D poses, the steps continue iteratively until convergence. Then final 3D pose is retrieved by minimizing the projection error in solution. The relative positions of the limbs and generate pose model that can be used to control 3D motion model. Finally, activity recognition is deeper and more reliable approach with deep neural network alongside human poses for understanding of human activity.

3.2.1. Spatial-Temporal Information Analysis

Human action recognition remains a problem to efficiently represent spatiotemporal skeleton sequences. To produce efficient sequences for each human body, the main concept is to measure the minimum distance between the detected pose and OpenPose library poses across the frame.

Let J_a, J_b, J_c, J_d, J_e be subset of body parts $J = \{1, \ldots, 17\}$ such as $J_a = \{3,4,5\} \subset J, J_b = \{6,7,8\} \subset J, \ldots, J_e = \{15,16,17\} \subset J$

The distance of generic pose $p_i(t)$ and generic prototypes $v \in V_l$ is obtained by

$$d_{\bar{J}_*}(p_i(t), v) = \frac{1}{|\bar{J}_*|} \sum_{j \in \bar{J}_*} \left\|(x_j, y_j)_i - (x_j^\dagger, y_j^\dagger)\right\|_2 \quad (2)$$

where $(x_j^\dagger, y_j^\dagger)$ are coordinates of j^{th} landmarks of v and $\bar{J}_*$ denote without missing landmark.

V_l be library of prototypes for action l . The embedding sequence is taking as:

$$D_{V_l, J_*}(t) = \min_{v \in V_l} d_{\bar{J}_*}(p_i(t), v) \quad (3)$$

Given a set of action $\mathcal{L}$, the meaningful sequences extracted from $p_i(t)$ is defined as :

$$Seq_i(V_l) = \left\{ D_{V_l, J}(t), D_{V_l, J_a}(t), \ldots, D_{V_l, J_e}(t) \right\} \forall l \in \mathcal{L} \quad (4)$$

3.3. Activity Recognition on Deep Neural Network

The system is developed process 2D data to 3D poses for human activity recognition. Human pose is basically represented as a graph where the joints are the nodes and the bones are the edges. Since every joint is connected to another joints in its neighbourhood and it has distribution power. The encoded structure matrix and model representation ability of weight matrix which become the dependent features in graph modelling.

3.3.1. Graph Modeling

To get better results with deeper network, VGG net is used in this system. VGG net is a Deep Neural Network architecture for object recognition. VGG net is designed as the simplest with 3x3 convolution and 2x2 max pooling layers are used throughout the whole network. Each layer along with pre-trained set of weights. It performs better on the training set because smaller and smaller features processing on additional layers. The training error actually decreases as the network gets deeper and can gain accuracy from considerably increased depth. By developing The novel adaptive dependency matrix and learn it through node embedding, our model can precisely capture the hidden spatial dependency and can generate a heat-map to encode a per-pixel likelihood for human joint localization. These heat maps are combined with temporal information alongside spatial information. Then, full-body information of 3D skeleton human pose is produced by simultaneously taking joint distribution in fully connected layers of Deep Neural Network.

3.3.2. Training and Testing

All through training, by taking the input from each sequence with rectangular patch around it and is resized to 224×224 RGB pixels. For data normalization, channel-wise RGB mean values are computed and subtracted from the images. The training labels to be regressed are multi-channel heat maps with each channel corresponding to the image location for each joint and is defined by using associated probability. The network consists of 3×3 convolutional layers with filters followed by ReLU activition function to provide dense prediction for all joints. A 2×2 max pooling layer is inserted after each of the first three convolutional layers and returns a set of corresponding probability to the class labels as output. The network is trained by minimizing the l_2 loss between the prediction and open source Caffe framework.

During testing, consistent with previous 3D pose methods, the subject with bounding box is assumed and the image patch in the bounding box is cropped in frame t and fed forward through the network to predict the probability of the joint occurring at each pixel. Finally, recognize the efficient understanding of human activity.

4. Experimental Result

4.1. Prepare Implementation

The deep neural network is trained by using Adam optimization algorithm and it takes advantage of momentum with moving part, the pre-trained weight model from Deep Learning framework and the start learning rate of 0.001 using 128 mini batch-size.

Fig. 2: Example result of our test on HMDB51 dataset in outdoor area

4.2. Result Analysis and Discussion

To describe the efficient results, we experiment on HMDB51dataset which contains the full complexity of video clips commonly found in YouTube, Google videos, movies and public database. As a result of our test in Fig.2. and the system shows well-defined to unseen activity in human movement and outperform good recognition. The proposed approach obtained high accuracy rates and the confusion matrix with colormap also states that the result is robust to multi-view as expressed in Fig.3. The accuracy result of training and testing as described in Fig.4.

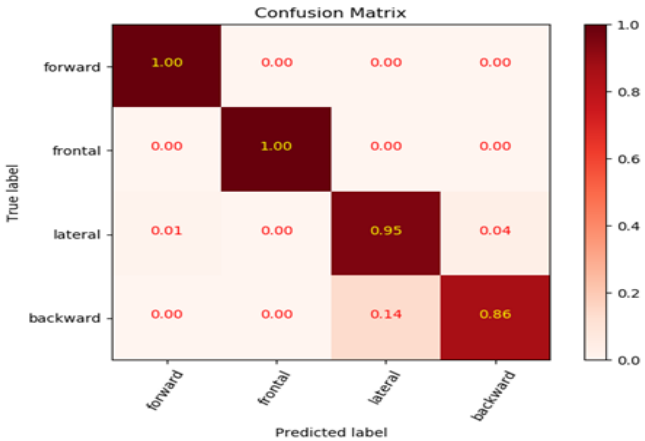


Fig. 3: The result of multi-view human recognition on confusion matrix

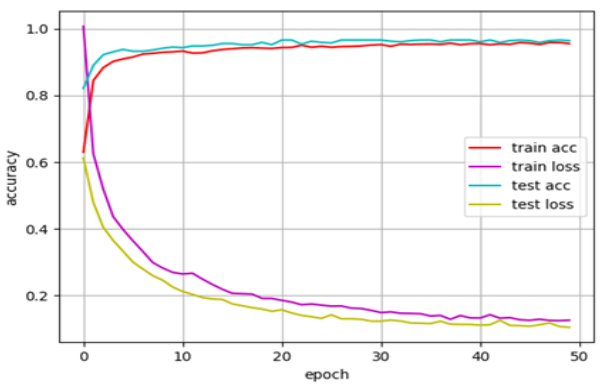


Fig. 4: The accuracy result of human recognition on training and testing

5. Conclusion and Future Work

There are various methods have been developed for human recognition by using skeleton sequence. But efficient multi-view human activity recognition systems have been required. That can record useful information and analyse the environment in many scopes. There are many challenges that are concerned with different variation in human pose, shape, illumination changes and background appearance. In this paper, the system is implemented by using deep neural network framework to get high accuracy recognition of human movement in indoor and outdoor areas. The experimental results have been concluded that recognition of moving object have a big dependency with different backgrounds, camera calibration and illumination changes. We trained and tested on HMDB51video dataset with different changes that are significantly increased recognition rate of our results.

Future research directions will continue 3D skeletal model for moving object with various dataset containing different activities and different views to describe the more accurate result of human activity recognition system on various data.

6. References

[1] J. Aggarwal, L. Xia, Human activity recognition from 3D data: A review, Pattern Recognition Letters 48 (0) (2014) 70–80.

[2] J. Han, L. Shao, D. Xu, J. Shotton, Enhanced computer vision with Microsoft Kinect sensor: A review, IEEE Transactions on Cybernetics 43 (5) (2013) 1318–1334.

[3] A. L. Brooks, A. Czarowicz, Markerless motion tracking: MS Kinect & Organic Motion OpenStage R, in: International Conference on Disability, Virtual Reality and Associated Technologies, 2012.

[4] S.Y. Jin, H.J. Choi, Essential body-joint and atomic action detection for human activity recognition using longest common subsequence algorithm, in: Workshops on Asian Conference on Computer Vision, 2013.

[5] D. Wu, L. Shao, Leveraging hierarchical parametric networks for skeletal joints action segmentation and recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014.

[6] D. C. Luvizon et al., 2D/3D Pose Estimation and Action Recognition using Multitask Deep Learning, IEEE Conference on Computer Vision Foundation, 2014.

[7] U. Iqbal, A. Doering, H. Yasin, B. Krüger, A. Weber, and J. Gall, A Dual-Source Approach for 3D Human Pose Estimation from a Single Image, 2017.

[8] W. Du, Y. Wang, and Y. Qiao, RPAN: An End-to-End Recurrent Pose-Attention Network for Action Recognition in Videos, in IEEE Int. Conf. on Computer Vision (ICCV), Oct. 2017, pp. 3745–3754.

[9] F. Yang, S. Sakti, Y. Wu, S. Nakamura, Make Skeleton-based Action Recognition Model Smaller, Faster and Better, in arXiv: 2019.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in CVPR, 2016, pp. 770–778.

[11] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields, in CVPR, 2017, pp.7291–7299.

ISBN 978-981-14-4787-7
Proceedings of 2020 the 10th International Workshop on Computer Science and Engineering (WCSE 2020)
Yangon (Rangoon), Myanmar (Burma), February 26- February 28, 2020, pp. 84-89

Compressed Sensing Image De-noising Algorithm Based on L1-L2 Norm Regularization

Liu Ziming[1], Fang Changjie[2+]

[1]School of Automation, Chongqing University of Posts and Telecommunications, China
[2]School of Science, Chongqing University of Posts and Telecommunications, China

Abstract. In this paper, we propose a compressed sensing image de-noising algorithm based on L1-L2 norm regularization. After the image is decomposed by the total variation spectral framework, L1 norm regularization is performed on the texture image, and L2 norm regularization is performed on the contour image, then the alternating direction method of multipliers (ADMM) is used for solution. The results of numerical experiment show that the proposed algorithm obtains higher peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) than the compared compressed sensing algorithm and the total variation algorithm, and can effectively maintain the contour information and texture information of the image when de-noising.

Keywords: compressed sensing, regularization, ADMM, total variation.

1. Introduction

Image noise is a problem in the field of image processing, and the noisy image will bring great difficulties to the feature extraction of the in the later stage[1].Therefore, the effective image de-noising has been the focus of attention. The purpose of image de-noising is to recover unknown original picture from the noisy image .In 1992, Osher and Fatemi proposed a classical total variation de-noising model [2]. However, the traditional total variation algorithm will misjudge the noise as the edge of the image, leading to staircase effect, which makes the quality of de-noised image unsatisfactory. On the other hand, the improved compressed sensing algorithm[3] based on the traditional total variation model can quickly reconstruct and de-noise the image, but the texture and structural features of the image are not taken into account.

In order to overcome this difficulty, based on the compressed sensing algorithm of reference [4], we propose a compressed sensing de-noising algorithm based on L1-L2 regularization. First, the noisy image is decomposed by the spectral framework to obtain a contour image with a small amount of noise and a texture image with a lot of noise. Since the contour image is mainly a smooth region, and is low-frequency information; while the texture image mainly contains details and noise, and is a high-frequency region [5], therefore, inspired by the reference[6], different weighting methods are used for the contour image and the texture image according to the structural features of the image. The L2 weighting method is adopted for the contour image with a small amount of noise, and the L1 weighting is adopted for the texture image with a large amount of noise, which can help to avoid the staircase effect in de-noising. Then, by combining the two parts after decomposition with the compressed sensing algorithm, a complex model with two regular terms is obtained. In order to solve this complex model, the alternating direction method of multipliers [7] (ADMM) similar to that in reference [8] is used for iterative solution. However, different from reference [8], in this paper, the Barziilai-Borwein method [9] in the gradient descent algorithm is directly adopted to solve the first sub-problem of the ADMM algorithm. Finally, the experiment results demonstrate that the proposed

[+] Corresponding author. Tel.: +86 13594142998
E-mail address: fangcj@cqupt.edu.cn

method is effective, and the objective PSNR and SSIM are superior to those of other algorithms of compressed sensing reconstruction.

2. Theoretical Foundation

The classical total variation regularization model [2] is as follows:

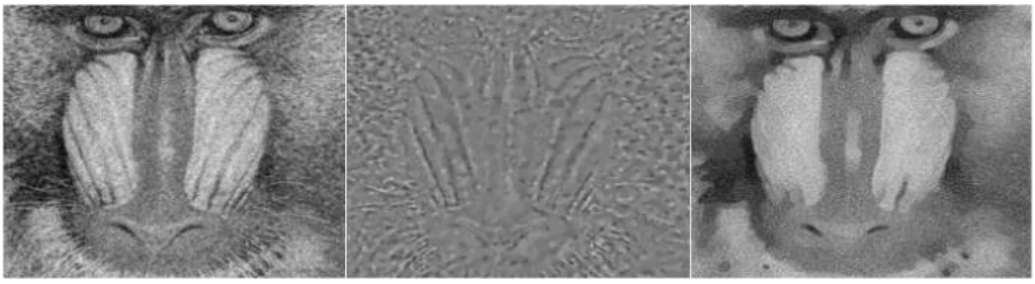
$$u = \arg \min_u \left(\frac{\kappa}{2}\int_\Omega |u-f|^2 du + \int_\Omega |Du| du\right) \tag{1}$$

where u is the original image, f represents noisy image, D is the gradient operator of forward difference, Ω is the domain of the image, κ is the parameter. The first term is a fitting term to ensure that the de-noised image is close to the original image; the second term is a regular term, which is also known as the total variation of u, and can promote the smoothness of the de-noised image.

The high-dimensional original value u is recovered from the low-dimensional observed value $f = Au + b$, where $f \in R^M$, $u \in R^N$, $A \in R^{M*N}$ ($M \ll N$) is the observation matrix. This kind of reconstruction problem is essentially an ill-posed inverse problem, which can be solved by optimization algorithm. Therefore, the total variation model combined with compressed sensing is as follows:

$$u = \arg \min_u \left(\frac{\kappa}{2}\int_\Omega |Au-f|^2 du + \int_\Omega |Du| du\right) \tag{2}$$

The total variation spectral framework, which was proposed by Gilboa, can decompose the original image into contour image and texture image [10, 11]. The image u can be decomposed into two parts: contour u_L and texture u_H.



(a) original image (b) contour image (c) texture image

Fig. 1: Images obtained by the decomposition of noisy image

3. Proposed Algorithm

The compressed sensing de-noising algorithm based on L1-L2 norm regularization is improved based on the compressed sensing algorithm proposed in reference [4], and it combines the advantages of weighted regularization and compressed sensing. In order to maintain the texture and edge information of the image while avoiding the staircase effect, the decomposed images are weighted and regularized:

$$\rho(u) = w_H \|Du_H\|_1 + w_L \|Du_L\|_2 \tag{3}$$

where $w_H = \eta/(1+\eta)$, $w_L = 1/(1+\eta)$, η is the absolute value of the image gradient mean.

The compressed sensing de-noising algorithm model based on L1-L2 norm regularization is:

$$(u_H, u_L) = \arg \min_{(u_H, u_L)} \left(w_H \|Du_H\|_1 + w_L \|Du_L\|_2 + \frac{\kappa}{2}\|Au-f\|_2^2\right) \tag{4}$$

The de-noising model (4) is a complex model with two regular terms. In order to facilitate the solution, ADMM is used to solve u_H and u_L respectively.

Fix u_L, the sub-problem of u_H can be obtained:

$$u_H = \arg \min_{u_H} \left(w_H \|Du_H\|_1 + \frac{\mu_H}{2}\|Au_H - f_H\|_2^2\right) \tag{5}$$

Fix u_H, the sub-problem of u_L can be obtained:

$$u_L = \arg \min_{u_L}(w_L \|Du_L\|_2 + \frac{\mu_L}{2}\|Au_L - f_L\|_2^2) \tag{6}$$

where $f_H = f - Au_L$, $f_L = f - Au_H$, μ_H and μ_L are the parameters. Obviously the problem after decomposition is easier to be solved than the original problem (4).

It is difficult to solve the sub-problem (5) directly. Therefore, the variable splitting [12] can be used to introduce the auxiliary variables u_1 and d_H. Set $u_1 = Au_H$ and $d_H = Du_H$. Then the sub-problem (5) can be converted to a constrained optimization problem.

$$\begin{aligned} &u_H = \arg \min_{u_H}(w_H \|Du_H\|_1 + \frac{\mu_H}{2}\|Au_H - f_1\|_2^2) \\ &s.t.\ \ u_1 = Au_H \quad d_H = Du_H \end{aligned} \tag{7}$$

The problem (7) can be converted to an unconstrained optimization problem using the Lagrange Multiplier Method, and its corresponding Lagrangian function is as follows:

$$\begin{aligned} &L_1(u_H, d_H, u_1; \gamma_H, \delta_H; \mu) = \|d_H\|_1 + \frac{\mu_H}{2w_H}\|u_1 - f_H\|_2^2 + \frac{\beta_{H1}}{2}\|Au_H - u_1\|_2^2 \\ &+ \frac{\beta_{H2}}{2}\|Du_H - d_H\|_2^2 - \gamma_H^T(Au_H - u_1) - \delta_H^T(Du_H - d_H) \end{aligned} \tag{8}$$

where β_{H1} and β_{H2} are penalty term coefficients, γ_H and δ_H are augmented Lagrangian coefficient matrix.

Since equation (8) contains three variables: u_H, d_H and u_1, it is complicated to solve. The ADMM can be used again to decompose it into three sub-problems with only one variable. Given u_H^K, d_H^K, u_1^k, γ_H^k, δ_H^k and μ_H^k ,we obtain

$$\begin{cases} u_H{}^{k+1} = \arg \min_{u_H} L_1(u_H, d_H^k, u_1^k; \gamma_H^k, \delta_H^k; \mu_H^k) \\ d_H{}^{k+1} = \arg \min_{d_H} L_1(u_H^{k+1}, d, u_1^k; \gamma_H^k, \delta_H^k; \mu_H^k) \\ u_1{}^{k+1} = \arg \min_{u_1} L_1(u_H^{k+1}, d_H^{k+1}, u; \gamma_H^k, \delta_H^k; \mu_H^{k+1}) \\ \gamma_H^{k+1} = \gamma_H^k - \beta_{H1}(u_1^{k+1} - Au_H^{k+1}) \\ \delta_H^{k+1} = \delta_H^k - \beta_{H2}(d_H^{k+1} - Du_H^{k+1}) \end{cases} \tag{9}$$

1) Fix d_H^K and u_1^k to solve u_H^{k+1}, u_H can be expressed as:

$$\begin{aligned} u_H = \arg \min_{u_H}(&\frac{\beta_{H1}}{2}\|Au_H - u_1\|_2^2 + \frac{\beta_{H2}}{2}\|Du_H - d_H\|_2^2 \\ &- \gamma_H^T(Au_H - u_1) - \delta_H^T(Du_H - d_H)) \end{aligned} \tag{10}$$

Gradient descent can be adopted to solve u_H, $u_H^{k+1} = u_H^k - \alpha_H^k g_H^k$, where g_H^k can be obtained by derivation, and α_H^k can be obtained by Barzilai-Borwein method [9].

$$\begin{aligned} &g_H^k = \beta_{H1}A^T(Au_H - u_1^k) + \beta_{H2}D^T(Du_H - d_H^k) - A^T\gamma_H^k - D^T\delta_H^k \\ &\alpha_H^k = \frac{v_H^{k\,T} v_H^k}{v_H^{k\,T} y_H^k} = \frac{(u_H^k - u_H^{k-1})^T(u_H^k - u_H^{k-1})}{(u_H^k - u_H^{k-1})^T[g_H^k(u_H^k) - g_H^k(u_H^{k-1})]} \end{aligned} \tag{11}$$

2)Fix u_H^{k+1} and u_1^k to solve d_H^{k+1}, d_H can be described as :

$$d_H = \arg \min_{d_H}(\|d_H\|_1 + (\beta_{H2}/2)\|Du_H - d_H + \delta_H^k/\beta_{H2}\|_2^2) \tag{12}$$

Equation (12) can be solved by the Shrinkage method [13]:

$$d_H^{k+1} = \max(\|Du_H^{k+1} + \delta_H^k/\beta_{H2}\|_1 - 1/\beta_{H2}, 0) * (Du_H^{k+1} + \delta_H^k/\beta_{H2}) / \|Du_H^{k+1} + \delta_H^k/\beta_{H2}\|_1 \tag{13}$$

3)Fix u_H^{k+1} and d_H^{k+1} to solve u_1^{k+1}, u_1 can be expressed as:

$$\min_{u_1}[(\mu_H / 2w_H)\|u_1 - f_H\|_2^2 + (\beta_{H1} / 2)\|u_1 - (Au_H + \gamma_H / \beta_{H1})\|_2^2] \tag{14}$$

Set the derivation of equation (14) equals to zero, then

$$u_1^{k+1} = \frac{\beta_{H1}(Au_H^{k+1} + \frac{\gamma_H^{k+1}}{\beta_{H1}}) + f_H \mu_H^{k+1} / w_H}{\beta_{H1} + \mu_H^{k+1} / w_H} \tag{15}$$

where the update of μ needs to meet the deviation criteria[14]. Set $\theta_H = \tau(\sigma_H^{k+1})^2 MN$, where σ_H^{k+1} is the noise standard deviation of f_H^k, MN is the total number of pixels in the image. If $\|u_1^{k+1} - f_H^k\| < c_H^{k+1}$, then μ_H is approaching 0. If $\|u_1^{k+1} - f_H^k\| > c_H^{k+1}$, we set $c_H^{k+1} = \|u_1^{k+1} - f_H^k\|$ and hence we have:

$$\mu_H^{k+1} = (\beta_{H1} / \sqrt{c_H^{k+1}})(\|Au_H^{k+1} + \gamma_H^{k+1} / \beta_{H1} - f_H\|_2^2 - 1 \tag{16}$$

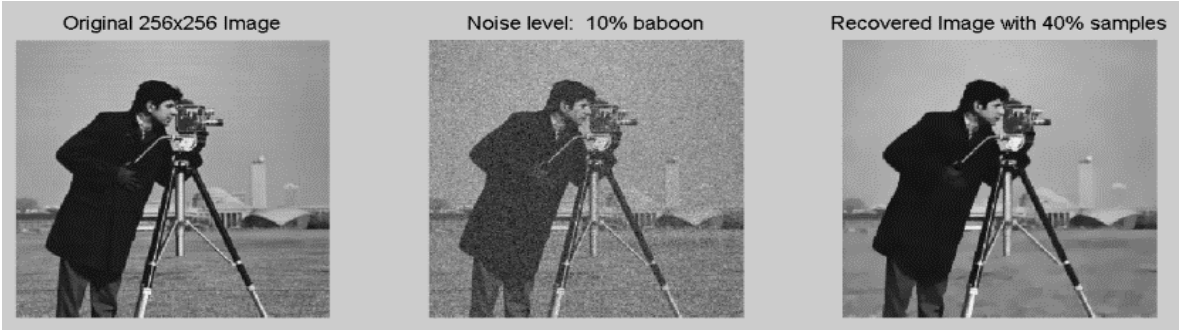
Similarly, two auxiliary variables u_2 and d_L are also needed to solve the sub-problem of u_L. Set $u_2 = Au_L$ and $d_L = Du_L$. In solving the sub-problem of u_L, the methods for solving u_L^{k+1} and u_2^{k+1} are similar to those for solving u_H^{k+1} and u_1^{k+1}. In other words, u_L^{k+1} and u_2^{k+1} can be obtained from u_H^{k+1} and u_1^{k+1} by replacing the corresponding subscript H with L, respectively. At the same time, d_L^{k+1} can be obtained in the following way different from d_H^{k+1}; see (17) below.Given u_L^k, d_L^k, u_2^k, γ_L^k, δ_L^k and μ_L^k, and fix u_L^{k+1} and u_2^k, d_L^{k+1} can be expressed as :

$$d_L^{k+1} = \max\left(\sqrt{S^{k+1} * \overline{S}^{k+1}} - \frac{1}{\beta_{L2}}, 0\right) * \frac{1}{\sqrt{S^{k+1} * \overline{S}^{k+1}} - \frac{1}{\beta_{L2}}} \tag{17}$$

where $S^{k+1} = Du_L^{k+1} - \delta_L^k / \beta_{L2}$, and $\overline{S}$ denotes the conjugate of S.

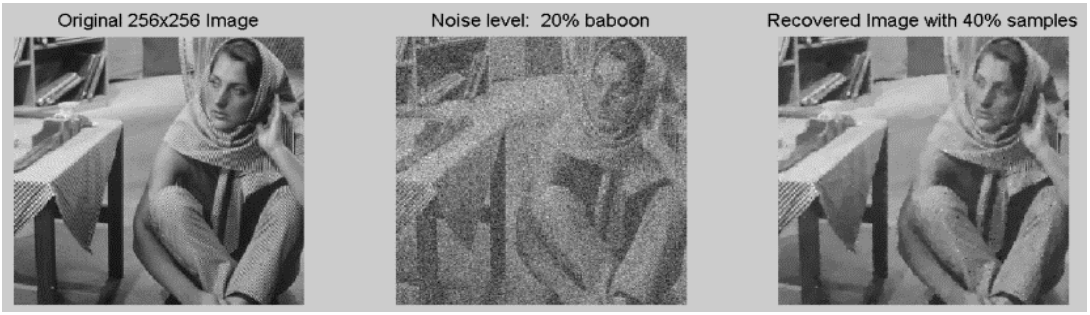
4. Experimental Results and Conclusion

Four classical digital images of 256 × 256 pixels: Lena, Cameraman, Barbara and Baboon are selected for the experiment. All the experiments were performed on MATLAB R2010b 7.11. The computer processor used in the experiment is Intel(R) i5-7300HQ, and RAM is 8.00GB. The noise is the Gaussian white noise with standard deviations of 0.05, 0.1, 0.2, and 0.3 respectively. The maximum number of iteration is set to 500; the algorithm precision tol is set as 10^{-4}, that is, the algorithm stops when $\|u^{k+1} - u^k\| / \|u^k\| < tol$. The penalty term parameters are set based on the experimental experience as $\beta_{H1} = 30$, $\beta_{H2} = 2*10^3$, $\beta_{L1} = 80$, $\beta_{L2} = 5$. SNR and SSIM are used as evaluation indexes for image reconstruction. The larger the index values of the two evaluations, the better. The unit of PSNR is dB, and the range of SSIM is from 0 to 1.The contrast experiment algorithms used in this work are the TV algorithm in reference[2] and compressed sensing algorithm in reference[4]. The compressed sensing algorithm and the proposed algorithm all use discrete wavelet transform (DWT) to sparse the noisy image. Gaussian random matrix is used for measurement matrix A, and the sampling ratio is 0.4.



(a) original image (b) noisy image (c) recovered image

Fig. 2: The image of cameraman de-noised by the proposed algorithm when the noise is 0.1.



(a) original image (b) noisy image (c) recovered image

Fig. 3: The image of barbara de-noised by the proposed algorithm when the noise is 0.2.

Table 1: Results of three de-noising algorithms

	Lena						Cameraman					
	TV		Reference[4]		Proposed algorithm		TV		Reference[4]		Proposed algorithm	
Noise	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
0.05	30.56	0.69	31.81	0.70	**32.94**	**0.72**	26.27	0.69	27.39	0.69	**28.5**	**0.72**
0.10	30.02	0.66	31.43	0.68	**32.85**	**0.70**	25.93	0.66	26.98	0.67	**28.04**	**0.69**
0.20	28.34	0.62	30.65	0.67	**32.50**	**0.68**	24.75	0.55	26.52	0.63	**28.01**	**0.65**
0.30	26.45	0.51	29.35	0.55	**31.28**	**0.63**	23.27	0.45	25.64	0.57	**27.57**	**0.63**
	Barbara						Baboon					
	TV		Reference[4]		Proposed		TV		Reference[4]		Proposed	
Noise	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
0.05	23.65	0.61	26.51	0.68	**27.96**	**0.70**	27.75	0.65	29.96	0.69	**31.84**	**0.79**
0.10	23.52	0.59	26.23	0.66	**27.79**	**0.68**	27.54	0.64	29.88	0.66	**31.42**	**0.77**
0.20	22.99	0.57	25.84	0.61	**27.66**	**0.66**	26.80	0.62	29.31	0.62	**30.73**	**0.71**
0.30	22.16	0.53	25.08	0.57	**26.54**	**0.59**	25.76	0.59	28.41	0.61	**29.70**	**0.67**

It can be seen from Table 1 that for the de-noising effect of the four test images, the value of PSNR is more than 1 dB compared to the contrast experiment algorithms. Especially when the image noise is large, the SSIM value of the proposed algorithm is higher, indicating that the de-noised image has a good structural similarity with the original image. This is because different regularization is adopted for the texture and contour image according to the different features during the de-noising process, which allows better texture features of the image after de-noising.

The proposed algorithm is a compressed sensing de-noising algorithm based on the L1-L2 norm regularization. Firstly, the image is decomposed into a contour image and a texture image by the total variation spectral framework method, and then the contour image and the texture image are regularized according to the image structural features. Then, ADMM is adopted for solution. Finally, the experimental results demonstrate that the de-noising effect of the proposed algorithm is better than that of the contrast experiment algorithms. However, it is worth noting that the gradient descent algorithm used for solving sub-problems has a large number of iterative steps. In future research, the conjugate gradient algorithm with fewer iteration steps can be considered.

5. References

[1] W. Wang, and C. J. He. A fast and effective algorithm for a poisson de-noising model with total variation. *IEEE Signal Processing Letters*. 2017, 24(3): 269-273.

[2] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physical D*. 1992, 60(1-4): 259-268.

[3] J. Wang, and C. L. Yang. An improved total variation model for image de-noising based on compression perception. *Computer Digital Eng*. 2007, 45(9): 1833-1836.(in Chinese)

[4] Z. P. Liu, and Y. Y. Chen. Total variation image de-noising algorithm base on compressed sensing. *Test and Measurement Technology*. 2018, 130(4): 53-58. (in Chinese)

[5] X. Liu, and L. Huang. A new nonlocal total variation regularization algorithm for image de-noising. *Mathematics Computers in Simulation*.2014, 97, 224-233.

[6] L. H. Yu, G. D. Liu, and C. J. Lin. An adaptive total variation de-noising algorithm for printed circuit board images. *Technology Eng. Sci*. 2019, 19(19): 207-213. (in Chinese)

[7] J. Yang, Y. Zhang, and W. Yin. A fast alternating direction method for tv l1-l2 signal reconstruction from partial fourier data. *IEEE Journal of Selected Topics in Signal Processing*. 2010, 4(2): 288-297.

[8] X. M. Yang. , Y. Q. Xiang, and Y. N. Liu. Image deblurring method with fraction-order total variation and adaptive regularization parameters. *Advanced Eng. Sci*. 2018, 50(6): 205-211. (in Chinese)

[9] J. Barzilai, and J. M. Borwein. Two-point step size gradient methods. *IMA Journal of Numerical Analysis*.1988, 8(1): 141-148.

[10] M. Chen, H. Zhang, and G. Lin, et al. A new local and nonlocal total variation regularization model for image de-noising. *Cluster Computing*.2019, 22: 7611-7627.

[11] Gilboa G. A total variation spectral framework for scale and texture. *Imaging Sci*.2014, 7(4): 1937-1961.

[12] Z. Qin, D. Goldfarn, and S. Ma. An alternating direction method for total variation de-noising. *Optimization Methods Software*.2015, 30(3): 594-615.

[13] Y. Wang, J. Yang, and W. Yin. A new alternating minimization algorithm for total variation image reconstruction. *International Conference on Wireless, Mobile and Multi-Media*, 2015.

[14] V. A. Morozov. Methods for solving incorrectly posed problems. *Springer-verlag Press*, 1984.

ISBN 978-981-14-4787-7
Proceedings of 2020 the 10th International Workshop on Computer Science and Engineering (WCSE 2020)
Yangon (Rangoon), Myanmar (Burma), February 26- February 28, 2020, pp. 90-94

QoE Comparison of AL-FEC Algorithms on H.265/HEVC Video and Audio Transmission with MMT

Toshiro Nunome[1+] and Koki Makino[2]

[1] Department of Computer Science, Graduate School of Engineering, Nagoya Institute of Technology, Japan
[2] Department of Computer Science, Faculty of Engineering, Nagoya Institute of Technology, Nagoya 466-8555, Japan

Abstract. This study assesses QoE of H.265/HEVC video and audio IP transmission over MMT with the two AL-FEC coding methods. We employ Reed-Solomon and Structured Low Density Parity Check. We perform multidimensional QoE assessment with three adjective pairs. We then find the appropriate selection of the coding method and the code rate according to the content and network condition can enhance QoE.

Keywords: MMT, streaming, QoE, AL-FEC, RS, S-LDPC

1. Introduction

MPEG has standardized MMT (MPEG Media Transport) [1] as an application-level media streaming protocol not only video but also various media types. It is standardized for replacing MPEG2-TS, which has been widely used in broadcasting, and for considering IP transmission. In the best-effort IP networks, packet loss and delay fluctuation occur, and then the output quality of media streaming can degrade. In MMT, we can utilize AL-FEC (Application Level Forward Error Correction) for recovering from errors and packet losses over the networks.

For network services, QoE (Quality of Experience) [2] is the ultimate quality metric. In [3], Nunome evaluates QoE of audio and video transmission with AL-FEC over MMT. The study does not utilize AL-FEC. Reference [4] introduces AL-FEC for QoE assessment of H.264/AVC video and audio transmission by means of MMT. The study employs Reed-Solomon as a coding algorithm. On the other hand, MMT has six candidates for coding algorithms [5].

Thus, this paper assesses QoE with two types of AL-FEC coding algorithms. We employ Reed-Solomon (RS) and Structured Low Density Parity Check (S-LDPC). S-LDPC has higher efficiency than RS for large data blocks. We then show the effect of the algorithms from a QoE point of view.

The remainder of this paper is organized as follows. Section 2 describes the experimental system and the QoE assessment method. Section 3 presents experimental results. Section 4 concludes this paper.

2. Experimental System

Figure 1 shows the experimental system. All the links in the network are 100 Mbps full-duplex Ethernet. Media Server transmits video and audio streams to Media Client through MMTP (MPEG Media Transport Protocol); it is an application-level protocol for multimedia transmission [1]. UDP is employed as the transport protocol under MMTP.

[+] Corresponding author. Tel.: +81-52-735-7785; fax: +81-52-735-5442.
E-mail address: nunome@nitech.ac.jp.

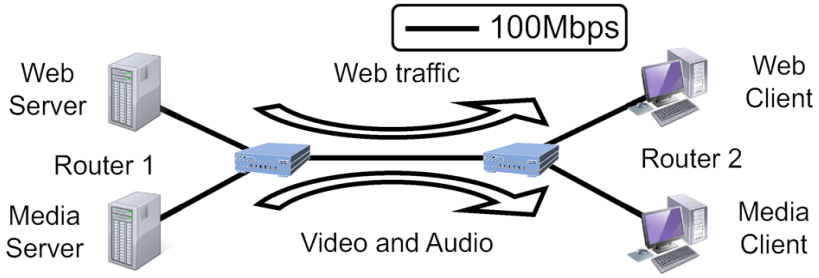


Fig. 1: Experimental network.

In audio, each MMTP/UDP packet includes an MU (Media Unit), which is an information unit for media synchronization control. For simplicity of implementation, we do not employ AL-FEC for audio.

As for the video, we consider a slice as a source packet for AL-FEC. The divided blocks from the slice are called source code blocks. From the blocks, the FEC code generation mechanism generates repair blocks. If the number of source code blocks is not enough to generate FEC repair blocks, we combine several slices and generate source code blocks from the combined slices. The size of the source code block in this paper is 512 bytes. This study is the first step; we then need to evaluate the effect of the source code block size in a future study.

We employ H.265/HEVC video (1920×1080 pixels) and AAC-LC (Advanced Audio Coding-Low Complexity) CBR (Constant BitRate) stereo audio. We utilize x265 as a video encoder. The video encoding bitrate is set to about 3 Mbps or 6 Mbps. We consider a video frame as a video MU. The MU rate is 29.97 MU/s. We deal with the picture pattern IPPPP (I+4P's) and the four slices per frame. The average bitrate and the MU rate of audio are 128 kbps and 46.875 MU/s, respectively. We employ two contents: *drama* (a scene of historical drama) and *sport* (a scene of figureskating). Both contents have BGM and scene changes, and sport has a larger movement than drama. The duration is 20 seconds.

Media Client outputs received audio and video after simple playout buffering control. We set the playout buffering time to 500 ms. In Media Client, we utilize FFmpeg for video decoding. Media Client does not use any error concealment techniques for video output.

As the interference traffic of audio and video, Web Server transmits Web traffic to Web Client according to requests generated by WebStone 2.5, which is a Web server benchmark tool. For the number of client processes, we employ 10 and 20.

To implement AL-FEC, we employ OpenFEC, which is an open-source library of FEC. We utilize Reed-Solomon and S-LDPC for the FEC algorithms. As the values of the code rate of FEC, we use 1/2, 2/3, and 5/6. We compare them with a method which does not perform FEC code generation (i.e., the code rate is 1). As the total video bitrate of the source video stream and FEC blocks, we consider two values: 3 Mbps and 6 Mbps. When we do not employ FEC, we use 3 Mbps or 6 Mbps as the video encoding bitrate. For employing FEC, we set the video encoding bitrate with consideration of the code rate: 1.5 Mbps and 3 Mbps for the code rate 1/2, 2 Mbps and 4 Mbps for the code rate 2/3, and 2.5 Mbps and 5 Mbps for the code rate 5/6.

For the QoE assessment, we perform a subjective experiment. In the experiment, we ask the assessors to evaluate video and audio output at Media Client. To reproduce the experimental conditions easily, we employ trace files which record the receive timing of video and audio MMTP packets. The assessors evaluate the first 10 seconds of the video and audio transmission.

The assessors are 15 male students of our university who major in computer science. We perform multidimensional QoE assessment with three adjective pairs shown in Table 1. Each adjective pair is scored within five grades. Score 5 means the right-side adjective of each pair, and score 1 represents the left-side one. Score 3 expresses moderate. We then calculate MOS (Mean Opinion Score) by averaging all the assessors' score for each adjective pair.

Table 1: Adjective pairs

item	adjective pair
video resolution	video is jaggy – clear
video collapse	video is corrupt – neat
overall quality	overall quality of video and audio is bad – excellent

3. Experimental Results

Before the consideration of QoE metrics, we mention the application-level QoS assessment. The application-level QoS is closely related to QoE because the application-level QoS is adjacent to QoE in the hierarchical network structure. We have evaluated the MU loss ratio (ratio of the number of MUs not output to the number of transmitted MUs), the slice loss ratio of output video (percentage of lost slices in an output frame), and the PSNR (Peak Signal-to-Noise Ratio) of output video. We then found that the application-level QoS is not largely affected by the AL-FEC coding algorithms. Hence, we focus on the QoE assessment results in this paper.

Figure 2 depicts the MOS of "video is jaggy – clear," which represents the sharpness of the image. The result of "video is corrupt – neat" is shown in Fig. 3; it is a measure for blockiness. Figures 2 and 3 are the results for the video bitrate 3 Mbps. We present the results of "overall quality of video and audio is bad – excellent" in Fig. 4 for the video bitrate 3 Mbps and in Fig. 5 for the video bitrate 6 Mbps. The figures also show the 95 % confidence intervals.

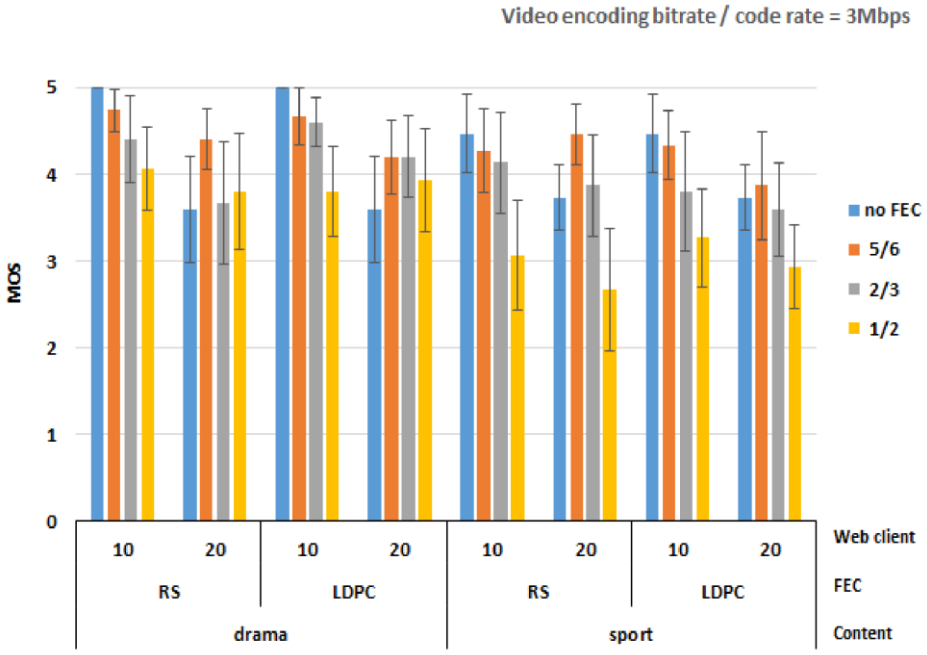


Fig. 2: MOS of "video is jaggy – clear" (video bitrate 3 Mbps).

We find in Fig. 2 that the methods without FEC have larger MOS values of "video is jaggy – clear" than those with FEC under the lightly loaded condition (10 Web clients). The main reason is the higher encoding bit rate in the methods without FEC. The higher encoding bit rate can provide better video image quality. On the other hand, the code rate 5/6 has the highest MOS value under the heavily loaded condition (20 Web clients). The corruption of video output makes the users difficult to notice a slight quality difference between 5/6 and without FEC, and then the effect of corruption becomes dominant for the users' quality perception.

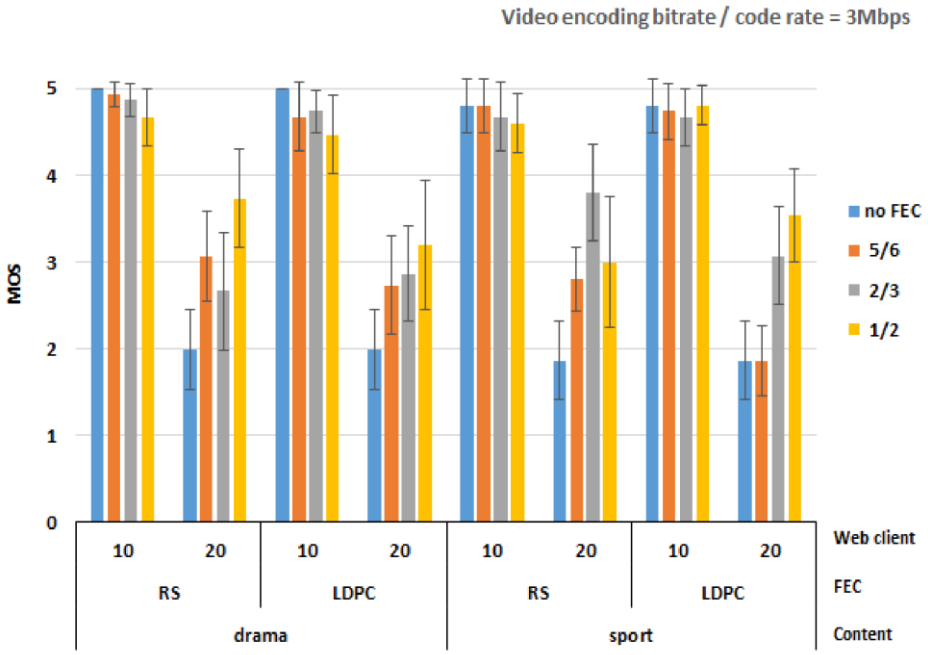


Fig. 3: MOS of "video is corrupt – neat" (video bitrate 3 Mbps).

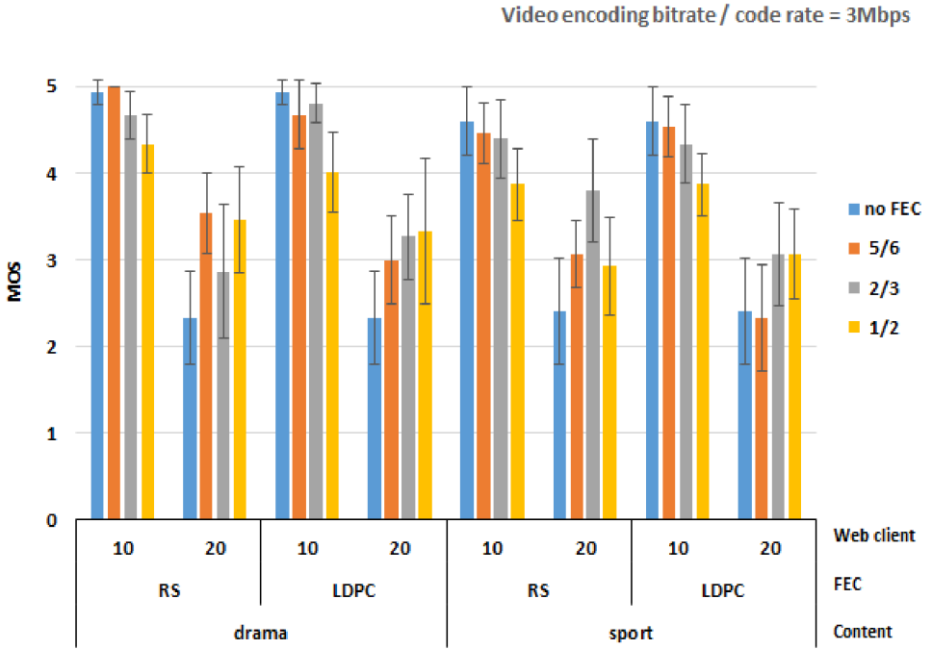


Fig. 4: MOS of "overall quality of video and audio is bad – excellent" (video bitrate 3 Mbps).

In Fig. 3, we hardly find the difference among the code rate values and the coding methods under the lightly loaded condition. This is because of the small loss ratio of video packets. On the other hand, on the heavily loaded condition, the MOS value of "video is corrupt – neat" is correlated with the code rate. The larger redundancy with the lower code rate can provide higher MOS value. For sport, the code rate 2/3 has the highest MOS value for RS, and the code rate 1/2 is the best for S-LDPC; i.e., the best code rate is different for the coding methods.

For the video bitrate 3 Mbps, we notice in Fig. 4 that the code rates 5/6 and 1/2 have the highest MOS values for RS and S-LDPC, respectively, for drama under the heavily loaded condition. Meanwhile, the code rate 2/3 is better for sport. This implies that a tradeoff relationship between FEC redundancy and encoding bitrate is affected by the contents and the coding algorithms.

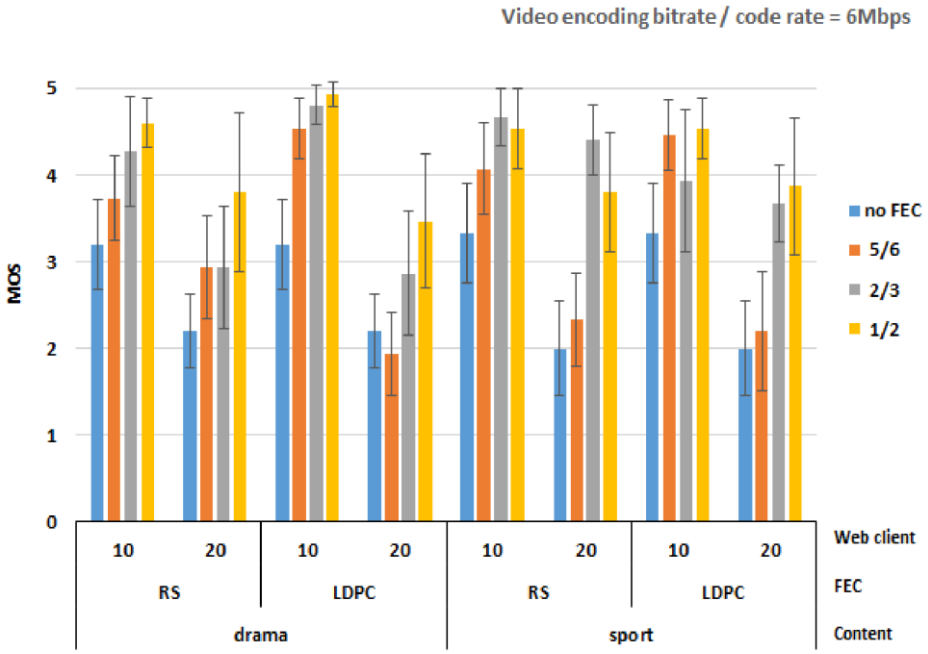


Fig. 5: MOS of "overall quality of video and audio is bad – excellent" (video bitrate 6 Mbps)

In Fig. 5, when the video bitrate is 6 Mbps, we see that the higher redundancy (lower code rate) has higher MOS value of "overall quality of video and audio is bad – excellent" for drama. This is because even the most redundant situation in this paper (i.e., the code rate 1/2), the video encoding bitrate is 3 Mbps; it can provide enough video quality.

4. Conclusions

In this study, we assessed the QoE of H.265/HEVC video and audio IP transmission over MMT with the two AL-FEC coding methods. We then found that the appropriate selection of the coding method and the code rate according to the content and network condition can enhance QoE.

In future work, we need to assess the effect of AL-FEC coding mechanisms under various conditions including wireless networks. We will also perform an evaluation with FEC mechanisms for audio.

5. References

[1] ITU-T Rec. P.10/G.100, "Vocabulary for performance, quality of service and quality of experience," Nov. 2017.

[2] ISO/IEC 23008-1, "Information technology – High efficiency coding and media delivery in heterogeneous environments – Part 1: MPEG media transport (MMT)," Second edition, Aug. 2017.

[3] T. Nunome, "A video output method for H.265/HEVC video and audio IP transmission and its QoE," Proc. 25th International Conference on Telecommunications (ICT 2018), pp. 259-263, June 2018.

[4] T. Nunome, "The joint effect of MMT AL-FEC and error concealment on video streaming QoE, " Proc. IEEE International Conference on Cloud Networking (CloudNet 2018), Oct. 2018.

[5] ISO/IEC 23008-10, "Information technology – High efficiency coding and media delivery in heterogeneous environments – Part 10: MPEG media transport forward error correction (FEC) codes," Apr. 2015.

Chapter 4
Communication and Information System

ISBN 978-981-14-4787-7
Proceedings of 2020 the 10th International Workshop on Computer Science and Engineering (WCSE 2020)
Yangon (Rangoon), Myanmar (Burma), February 26- February 28, 2020, pp. 95-102

Projected Iterative MVDR Beamforming for Null Broadening and First Sidelobe Suppression in the Presence of DOA Mismatch

Raungrong Suleesathira[+]

Department of Electronic and Telecommunication Engineering, Faculty of Engineering
King Mongkut's University of Technology Thonburi, Thung-khru, Bangmod, Bangkok, 10140 Thailand

Abstract. Although the minimum variance distortionless response (MVDR) beamformer can keep the desired signal from interfering signals, however, it has notch null and no sidelobe level control which may lead to performance degradation in the case of scattering environment or unexpected interfering signals such as rapidly moving jammer environment. It also has high sidelobe levels for either low number of samples or high input signal to noise ratio (SNR). The projected iterative MVDR (PI-MVDR) beamformer is presented to broaden the width of nulls, suppress the first sidelobe level and support the case of low number of samples no matter of low or high input SNR. The averaged covariance matrix used in the PI-MVDR beamformer can overcome the circumstance of low number of samples over any input SNR value. Projecting the averaged covariance matrix on to the interference space enables the iterative MVDR beamformer to steer a mainbeam toward the desired signal direction and make a broad null in the direction of an unwanted signal. To reduce the first sidelobe level, find the direction corresponding to the strongest sidelobe close to the mainbeam which is considered as a direction of arrival (DOA) of a hypothetical interference signal and employ the PI-MVDR beamformer to produce extra wide null. The steering vector of the desired signal is estimated to illustrate the robustness of the PI-MVDR beamformer in the presence of DOA mismatch. Simulation results demonstrated the achievement of the presented approach.

Keywords: beamformer, covariance matrix, direction of arrival

1. Introduction

Array beamforming methods make use of the spatial dimension to combat interference signals and noise by producing nulls at the direction of interference signals and a mainbeam at the direction of the desired signal. A variety of issues have been challenged to be solved. The performance can be severely degraded if the DOA is spread due to multipath propagation or rapidly moving interferences. Forming broad nulls around the directions of interferences can solve the problem since the broaden null can reject the interfering sources not only from a specific direction but from a specific spatial region as well. As a result, null broadening allows the interferer to move in a certain area. New interferences might exist during the time span processing. Suppressing the sidelobes especially the ones close to the mainbeam would be benefit to reject the new interferences disturbing the desired signal. Most beamformers dramatically degraded when received signal is either taken over few samples or strong, the weights might not only give a proper beampattern corresponding to the DOA of the received signal and high sidelobe as well. Another challenge is the sensitiveness to the mismatch between the actual and presumed steering vector of the desired signal.

Covariance matrix taper (CMT) is a classical approach of null broadening [1]. Adaptive variable diagonal loading combined with the CMT approach is presented in [2] for null broadening beamforming to develop the CMT. The CMT and projection are combined, the null depth is deeper than that of the CMT approach [3]. The researches that focus only on the sidelobe reduction have been conducted. Recently, an iterative beamforming is proposed in order to improve the three conventional beamformers by placing extra

[+] Corresponding author. Tel.: +6695257197
E-mail address: raungrong.sul@kmutt.ac.th

nulls to achieve the desired sidelobe level [4]. Efforts to broaden nulls and control the sidelobe levels have been done. In [5], the CMT is constructed to broaden the width of nulls for interference signal sources. Constraint of nulls and sidelobe levels are used to guarantee that the level is strictly lower than the prescribed threshold value. Certainly, the optimal solution needs complex computation. In [6], it uses the CMT to broaden the width of nulls and uses the support vector match regression to control the sidelobe level.

The rest of paper is organized as follows. The signal model is described in Section 2. The conventional MVDR beamforming is updated in Section 3. The averaged covariance matrix is given in Section 4. Section 5 presents an estimation of the steering vector of the desired signal used in the presence of DOA mismatch. The projected iterative MVDR beamformer is proposed in Section 6 for null broadening and first sidelobe suppression. The evaluation of the proposed method is verified in the simulation results of Section 7. Finally, conclusions are drawn in Section 8.

2. Signal Model

Consider a uniform linear array (ULA) comprising L omni-directional antenna elements with equispaced d as shown in Fig. 1. A narrowband source of interest and interferences are transmitted in the far field region of the array. The received signal at the sampling time k is mathematically represented as [7]

$$\mathbf{x}(k) = s_d(k)\mathbf{a}(\theta_d) + \sum_{i=1}^{I} s_i(k)\mathbf{a}(\theta_i) + \mathbf{n}(k) \tag{1}$$

The $L\times 1$ received signal vector $\mathbf{x}(k) = [x_1(k) \quad x_2(k) \quad \dots \quad x_L(k)]^T$ consists of the desired signal, a sum of I interference signals ($(I+1) < L$) and an additive noise. Let denote $s_d(k)$ be the desired signal waveform in the directions of arrival (DOA) of θ_d, $s_i(k)$ be the ith interference waveform in the DOA of θ_i. An $L\times 1$ steering vector is expressed as $\mathbf{a}(\theta) = [1 \quad e^{j2\pi\frac{d}{\lambda}\sin\theta} \quad \dots \quad e^{j2\pi(L-1)\frac{d}{\lambda}\sin\theta}]^T$ where λ is the signal wavelength, $(\cdot)^T$ denotes the transpose operation. The steering matrix with columns of the steering vector is given as $\mathbf{A} = [\mathbf{a}(\theta_d) \quad \mathbf{a}(\theta_1) \quad \dots \quad \mathbf{a}(\theta_I)]$. Then, the received signal in Eq. (1) can be rewritten as

$$\mathbf{x}(k) = \mathbf{A}\mathbf{s}(k) + \mathbf{n}(k) \tag{2}$$

where $\mathbf{s}(k) = [s_d(k) \quad s_1(k) \quad \dots \quad s_I(k)]^T$ is an $(I+1)\times 1$ vector of the desired signal and interferences impinging on the antenna array. Assume that the desired and interference signals are uncorrelated, i.e.,

$$E[s_d(k)s_d^*(k)] = p_d \ , \ E[s_d(k)s_i^*(k)] = 0 \text{ for } i = 1,\dots I \text{ and } E[s_i(k)s_j^*(k)] = \begin{cases} 0 & i \neq j \\ p_i & i = j \end{cases}$$

where p_d and p_i denotes the power of the desired signal and the ith interference signal, respectively. $\mathbf{n}(k) = [n_1(k) \quad n_2(k) \quad \dots \quad n_L(k)]^T$ is an $L\times 1$ complex noise vector with zero mean and variance σ_n^2. In fact, the directional sources and the noise are uncorrelated, i.e., $E[s_d(k)n_l^*(k)] = 0$ and $E[s_i(k)n_l^*(k)] = 0$ for $i = 1,\dots,I$ and $l = 1,\dots,L$. The noise on different elements is also assumed to be uncorrelated and independent, i.e. $E[n_j(k)n_l^*(k)] = E[n_j(k)]E[n_l^*(k)] = \begin{cases} 0 & j \neq l \\ \sigma_n^2 & j = l \end{cases}$ where $E[\bullet]$ denotes the expectation operator and $(\cdot)^*$ denotes the complex conjugate.

The covariance matrix of the received signal is formulated as $\mathbf{R}_x = E[\mathbf{x}(k)\mathbf{x}^H(k)]$. By using $\mathbf{x}(k)$ in Eq. (1), it becomes

$$\mathbf{R}_x = p_d\mathbf{a}(\theta_d)\mathbf{a}^H(\theta_d) + \sum_{i=1}^{I} p_i\mathbf{a}(\theta_i)\mathbf{a}^H(\theta_i) + \sigma_n^2\mathbf{I}_L \tag{3}$$

where $\mathbf{I}_L$ is the $L \times L$ identity matrix and $(\cdot)^H$ denotes the complex conjugate transposition operation. Eq. (3) can be considered as a sum of the covariance matrix of the desired signal as $\mathbf{R}_d = p_d \mathbf{a}(\theta_d)\mathbf{a}^H(\theta_d)$ and the interference-plus-noise covariance matrix as $\mathbf{R}_{i+n} = \sum_{i=1}^{I} p_i \mathbf{a}(\theta_i)\mathbf{a}^H(\theta_i) + \sigma_n^2 \mathbf{I}_L$.

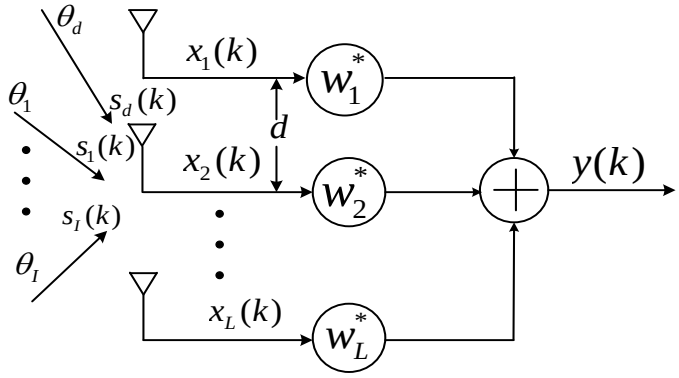


Fig. 1: A ULA of L antenna elements receiving $(I+1)$ directional sources

3. Iterative MVDR Beamformer

The output signal can be calculated by $y(k) = \mathbf{w}^H \mathbf{x}(k)$ where $\mathbf{w} = [w_1 \quad w_2 \quad \ldots \quad w_L]^T$ is an $L \times 1$ weight vector. Given $\mathbf{w}$, the output SINR is found by

$$\mathrm{SINR}(\mathbf{w}) = \frac{p_d \,|\, \mathbf{w}^H \mathbf{a}(\theta_d) \,|^2}{\mathbf{w}^H \mathbf{R}_{i+n} \mathbf{w}}. \tag{4}$$

Under assumption that the steering vector of the desired signal $\mathbf{a}(\theta_d)$ is known precisely, the minimum variance distortionless response (MVDR) beamformer is obtained by minimizing the denominator of Eq. (4), i.e., minimizing the variance/power of interference and noise while keeping the numerator of Eq. (4) fixed, i.e., ensuring the output desired signal without distortion towards the direction of the desired source. This is equivalent to a linearly constrained quadratic optimization problem as

$$\min_{\mathbf{w}} \mathbf{w}^H \mathbf{R}_{i+n} \mathbf{w} \quad \text{subject to} \quad \mathbf{w}^H \mathbf{a}(\theta_d) = 1 \ . \tag{5}$$

Using the Lagrange multiplier, the optimal solution for this weight vector is obtained as [8]

$$\mathbf{w}_{MDVR} = \alpha \mathbf{R}_{i+n}^{-1} \mathbf{a}(\theta_d) \tag{6}$$

where $\alpha = 1/\mathbf{a}^H(\theta_d)\mathbf{R}_{i+n}^{-1}\mathbf{a}(\theta_d)$ and $(\cdot)^{-1}$ denotes the matrix inversion. To avoid the inverse matrix calculation in Eq. (6), the solution can be obtained by weight adjustment as [8]

$$\mathbf{w}_{m+1} = \mathbf{w}_m - c(\mathbf{R}_{i+n}\mathbf{w}_m - \mathbf{a}(\theta_d)). \tag{7}$$

For fast convergence, set the constant $c = 2/(\lambda_{\min} + \lambda_{\max})$ where $\lambda_{\min}$ and $\lambda_{\max}$ are the minimum and maximum eigenvalues of $\mathbf{R}_{i+n}$, respectively. For some amount of number of iterations, the weight vector in Eq. (7) converges to $\mathbf{w}_{MDVR}$ [9-10].

4. Averaged Covariance Matrix

Since it is difficult to obtain $\mathbf{R}_{i+n}$ in practice, this matrix is commonly replaced by the sample covariance matrix of the received signal with limited number of snapshots K as $\hat{\mathbf{R}}_x = \frac{1}{K}\sum_{k=0}^{K-1} \mathbf{x}(k)\mathbf{x}^H(k)$.

When $K \to \infty$, $\hat{\mathbf{R}}_x$ will converge to the theoretical covariance matrix $\mathbf{R}_x$. There are two drawbacks of replacement. First, when K is small, a gap between $\hat{\mathbf{R}}_x$ and $\mathbf{R}_x$ becomes larger. The limited number of snapshots can cause the disturbance of small eigenvalues corresponding to the noise subspace. As a result, the performance of the beamforming will be degraded due to the rise of sidelobes. Secondly, the

performance degradation will be significant as input SNR increases which can be proven by substituting $\mathbf{R}_x = \mathbf{R}_d + \mathbf{R}_{i+n}$ into $\mathbf{w}^H \mathbf{R}_x \mathbf{w}$. Since $\mathbf{w}^H \mathbf{a}(\theta_d) = 1$, it yields as

$$\mathbf{w}^H \mathbf{R}_x \mathbf{w} = p_d + \mathbf{w}^H \mathbf{R}_{i+n} \mathbf{w}. \tag{8}$$

Thus, as the power p_d increases, such replacement can cause large error.

Consider the case of low number of samples $K = 30$ and high input SNR ($10\log_{10}(p_d / \sigma_n^2)$) 3 dB. Fig. 2 shows the comparison the optimal beamformer $\mathbf{w}_{MDVR}$ using the sample covariance matrix of the receive signal $\hat{\mathbf{R}}_x$ and the sample covariance matrix of the interference-plus-noise signal given by $\hat{\mathbf{R}}_{i+n} = \frac{1}{K}\sum_{k=0}^{K-1}(\mathbf{x}_i(k) + \mathbf{n}(k))(\mathbf{x}_i(k) + \mathbf{n}(k))^H$. A ULA of 10 elements is used with inter-element spacing $d = \lambda / 2$. The desired signal is taken at $\theta_d = 0°$ and two interferences are taken at $\theta_1 = -45°$ and $\theta_2 = 45°$. The beampattern using $\hat{\mathbf{R}}_x$ (blue line) gives the bad mainbeam and shallow nulls. For the beampattern using $\hat{\mathbf{R}}_{i+n}$ (red line), the mainbeam is unity at $\theta_d = 0°$ and the two notch nulls are at $\theta_1 = -45°$, $\theta_2 = 45°$. The result assures that the replacement $\mathbf{R}_{i+n}$ by $\hat{\mathbf{R}}_x$ degrades the performance of the MVDR beamformer if high input SNR is present in the sample data or low number of snapshot is available.

Then, an estimate of the interference-plus-noise covariance matrix is needed to achieve a better beampattern. The spatial spectrum distribution in all direction is first created as

$$P(\theta) = \frac{1}{\mathbf{a}^H(\theta)\mathbf{E}_{i+n}\mathbf{E}_{i+n}^H\mathbf{a}(\theta)} \tag{9}$$

where $\mathbf{E}_{i+n}$ is the interference plus noise subspace found by applying the singular value decomposition to $\hat{\mathbf{R}}_x$ as $\hat{\mathbf{R}}_x = \sum_{l=1}^{L} e_l \mathbf{e}_l \mathbf{e}_l^H$. Descend the order of eigenvalues e_l as $e_1 > e_2, \ldots, > e_L$. According to Eqs. (1) and (2), the received signal is the addition of source of interest, interferences and noise. Consequently, the interference plus noise subspace consists of the eigenvectors belong to the small eigenvalues built as $\mathbf{E}_{i+n} = [\mathbf{e}_2 \quad \mathbf{e}_3 \quad \ldots \quad \mathbf{e}_L]$. The interference-plus-noise covariance matrix can be reconstructed by utilizing the spectrum to integrate over a region separated from the desired directions [11].

$$\tilde{\mathbf{R}}_{i+n} = \int_{\bar{\Theta}} P(\theta)\mathbf{a}(\theta)\mathbf{a}^H(\theta)d\theta \tag{10}$$

where $\bar{\Theta}$ is the complement sector of Θ and Θ is an angular sector where the desired signal is located. $\Theta \cup \bar{\Theta}$ covers the whole spatial domain while $\Theta \cap \bar{\Theta}$ is empty.

However, at low input SNR, the replacement of $\mathbf{R}_{i+n}$ by $\hat{\mathbf{R}}_x$ results in a small error, since the desired signal power p_d is small. At high input SNR, we need an estimate instead of the replacement. In order to select the proper covariance matrix corresponding to the input SNR value, $\mathbf{R}_{i+n}$ will be replaced by the averaged covariance matrix $\mathbf{R}_{ave}$ as [12]

$$\mathbf{R}_{ave} = \alpha\tilde{\mathbf{R}}_{i+n} / \|\tilde{\mathbf{R}}_{i+n}\|_F + (1-\alpha)\hat{\mathbf{R}}_x / \|\hat{\mathbf{R}}_x\|_F \tag{11}$$

where $\|\cdot\|_F$ is the frobenius norm. The parameter α is defined as

$$\alpha = \frac{\mathbf{a}^H(\theta_d)\hat{\mathbf{R}}_x\mathbf{a}(\theta_d)}{\|\hat{\mathbf{R}}_x\|_F \|\mathbf{a}(\theta_d)\|^2} \tag{12}$$

where $\|\cdot\|$ is the norm of a vector. The value of α is between zero and one which can be used as an indicator of the desired signal power compared with the interference signal power. When α is equal to zero, it means that there is no signal from the direction θ_d, then $\hat{\mathbf{R}}_x$ can be used for low input SNR. When α is equal to one, it means that the eigenvector associated to the largest eigenvalue of $\hat{\mathbf{R}}_x$ is equal to $\mathbf{a}(\theta_d)$, then $\tilde{\mathbf{R}}_{i+n}$ should be used for high input SNR. Figure 3 shows the α values increase versus the input SNR.

5. Steering Vector of the Desired Signal Estimation

When the DOA of the desired signal mismatch exists, the performance of the MVDR beamformer and the iterative MVDR beamformer will be significantly degraded. Also, the value of α will reflect incorrect meanings. The mismatch between the actual DOA θ_d and the presumed DOA $\bar{\theta}_d$ falls in the uncertainty region defined as $\Theta = [\theta_d - \Delta_d, \theta_d + \Delta_d]$ where Δ_d is the angular sector of mismatch. Since only one desired signal assumed in the uncertainty region, the eigenvector with the largest eigenvalue of the constructed matrix will be the steering vector estimate. Build a matrix over the uncertainty region as $\mathbf{V} = \int_{\Theta} \frac{\mathbf{a}(\theta)\mathbf{a}^H(\theta)}{\mathbf{a}(\theta)\hat{\mathbf{R}}_x^{-1}\mathbf{a}^H(\theta)}$. Apply the singular value composition to the matrix $\mathbf{V}$ to obtain the most principle eigenvector $\mathbf{v}_{ds}$. The steering vector of the desired signal can be estimated as $\bar{\mathbf{a}}(\theta_d) = \frac{\mathbf{v}_{ds}}{|v_{ds}|}$ where $|v_{ds}|$ is the absolute value of the first element in the vector $\mathbf{v}_{ds}$ [13].

6. Projected Iterative MVDR Beamformer

For low input SNR, the weight vector $\mathbf{w}_{MVDR}$ is able to form a mainbeam in the look direction of the desired signal and nulls in the undesired directions belong to the DOAs of interference signals. However, the null is notch and it does not take into account the sidelobe levels. Since a moving interference is a serious problem in the antenna array, we then assume that the direction of the desired signal θ_d is constant whereas θ_i , on the other hand, is not constant. In a local time-varying scattering environment, the interference direction is randomly selected from the interval $(\theta_i - \Delta\theta/2, \theta_i - \Delta\theta/2)$ where $\Delta\theta$ is the total direction change so-called the angular spread which is used to determines the null width.

New interference signals may arise during the time span that a certain data sample set is processed. Since this sample set does not contain these new interference signals, then no null is create to cancel, thus resulting in SINR degradation. Especially, if these new interference signals have DOA close to the directions of the mainlobe which is the DOA of the desired signal. In such case, a low first sidelobe beside the mainlobe would help to keep the SINR at high levels. The nulls can be broadened and the sidelobe level can be controlled by the following projection transform [3].

The projection approach begins with constructing the correlation matrix of I interference signals as $\mathbf{Z}_P = \sum_{i=1}^{I} \mathbf{a}(\theta_i)\mathbf{a}^H(\theta_i)$. Next, the matrix $\mathbf{Z}_p$ is tapered by $\mathbf{R}_P = \mathbf{Z}_P \circ \mathbf{T}$ where $\circ$ represents Hadamard product and the sample of the *a*th row and *b*th column of the tapered matrix $\mathbf{T}$ is expressed as $T_{ab} = \frac{\sin((a-b)\Delta\theta)}{(a-b)\Delta\theta}$. Decompose the matrix $\mathbf{R}_P$ as $\mathbf{R}_P = \sum_{l=1}^{L} \upsilon_l \mathbf{v}_l \mathbf{v}_l^H$. By chosing J eigenvectors belonging to the J largest eigenvalues, the projection matrix is $\mathbf{T}_P = \sum_{l=1}^{J<L} \mathbf{v}_l \mathbf{v}_l^H$. Then, the projected covariance matrix can be modified as $\mathbf{R}_{PRO} = \mathbf{T}_P \mathbf{R}_{ave} \mathbf{T}_P^H$. The projected iterative MVDR (PI-MVDR) beamformer is obtained by replacing $\mathbf{R}_{i+n}$ with $\mathbf{R}_{PRO}$ as [9-10]

$$\mathbf{w}_{m+1} = \mathbf{w}_m - c(\mathbf{R}_{PRO}\mathbf{w}_m - \mathbf{a}(\theta_d)) \ . \tag{13}$$

Let $\mathbf{w}_{PI-MVDR}$ denotes the weight vector after the iteration is terminated.

After obtaining $\mathbf{w}_{PI-MVDR}$, the next step is to suppress the sidelobe levels. The proposed algorithm to suppress the first sidelobe levels is summarized by the flowchart drawn in Fig. 4.

7. Simulation Results

A 30-element ULA with a spacing of half wavelength is used. The additive noise is modeled as a complex Guassian spatially and temporally white process with zero mean. Two interfering sources

$(I=2)$ are from directions of $\theta_1=-45°$ and $\theta_2=45°$ with the angular spread $\Delta\theta=10°$. The input SNR is equal to 3 dB and the INR (interference to noise ratio) is equal to 6 dB. The number of snapshot is set to $K=30$. The actual DOA of the desired signal is $\theta_d=5°$. The angular sector of mismatch is $\Delta_d=2°$. The possible angular sector of the desired signal is located $\Theta=[3°\ 7°]$, so the complement sector is $\bar{\Theta}=[-90°\ 3°)\cup(7°\ 90°]$. The number of weight adjustment to terminate the iteration is 15.

Figure 5 shows the PI-MVDR beamformer by using $\hat{\mathbf{R}}_x$, $\hat{\mathbf{R}}_{i+n}$, $\tilde{\mathbf{R}}_{i+n}$ and $\mathbf{R}_{ave}$ to generate $\mathbf{R}_{PRO}$. The beam patterns using $\hat{\mathbf{R}}_x$, $\hat{\mathbf{R}}_{i+n}$ (blue and purple lines) is bad when the number of snapshot is few and the input SNR is high. The beampattern using $\tilde{\mathbf{R}}_{i+n}$ (black line) can reduce the sidelobe levels but still has shallow nulls since the interference and noise subspace generated is not good for low number of samples. The beampattern using $\mathbf{R}_{ave}$ (red line) provides not only the right mainbeam at the desired direction $\theta_d=5°$ and also broad nulls between $-50°$ to $-40°$ and $40°$ to $50°$ corresponding to the angular spread of each interference direction.

In Figure 6, the beampattern after suppression (red line) can reduce the first sidelobe levels as marked by the two ellipses and still keep the null width and the mainbeam unchanged. Also, this does not affect the position of the nulls initially placed to cancel the interference signals.

It can be observed in Figure 7 that the output SINR increases significantly especially in the negative range input SINR. Due to the high gain of the output SINR over the input SINR, it implies that the interference signals at $\theta_1=-45°$, $\theta_2=45°$ are mitigated efficiently.

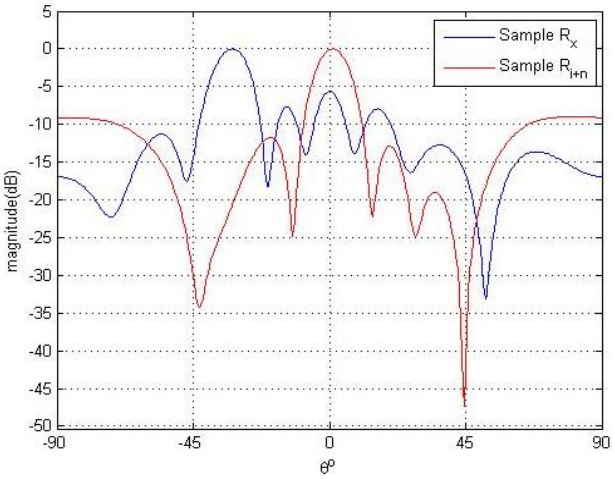


Fig. 2: Beampatterns by the MVDR Beamformer using $\hat{\mathbf{R}}_x$ and $\hat{\mathbf{R}}_{i+n}$

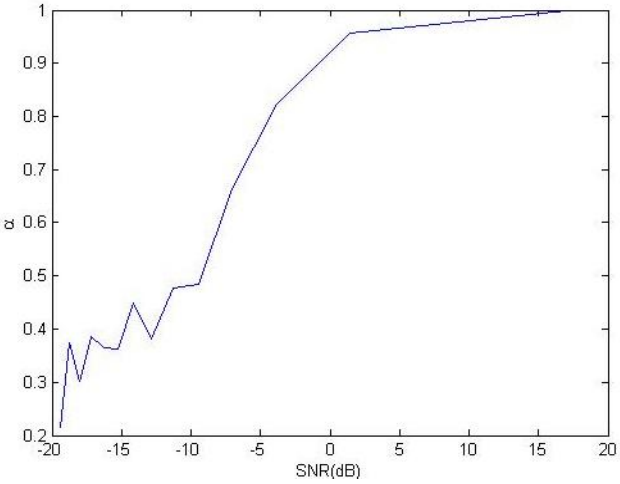


Fig. 3: The values of α versus the input SNR

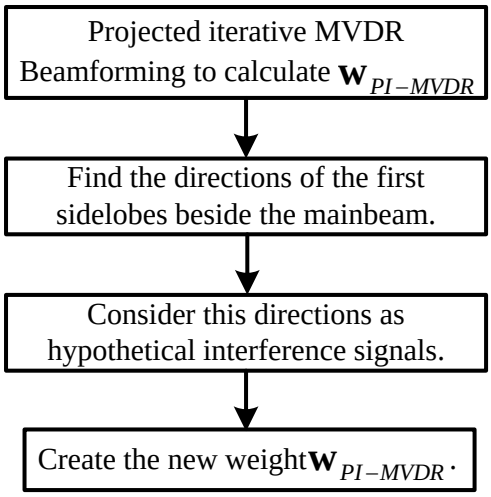


Fig. 4: Flowchart of the sidelobe suppression

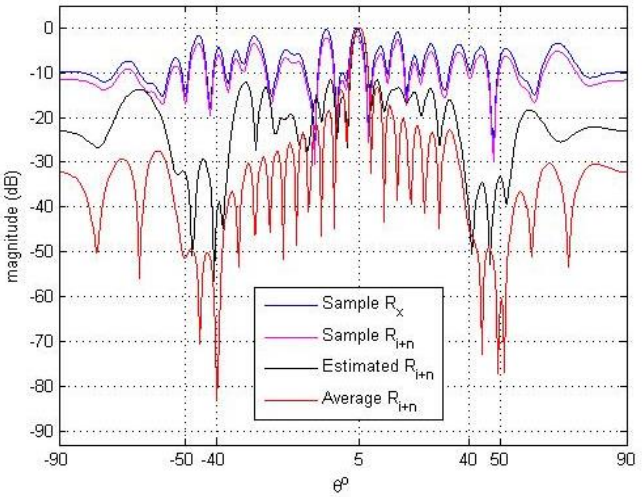


Fig. 5: Beampattrns by the PI-MVDR beamformer using $\hat{\mathbf{R}}_x$, $\hat{\mathbf{R}}_{i+n}$, $\tilde{\mathbf{R}}_{i+n}$ and $\mathbf{R}_{ave}$ to generate $\mathbf{R}_{PRO}$

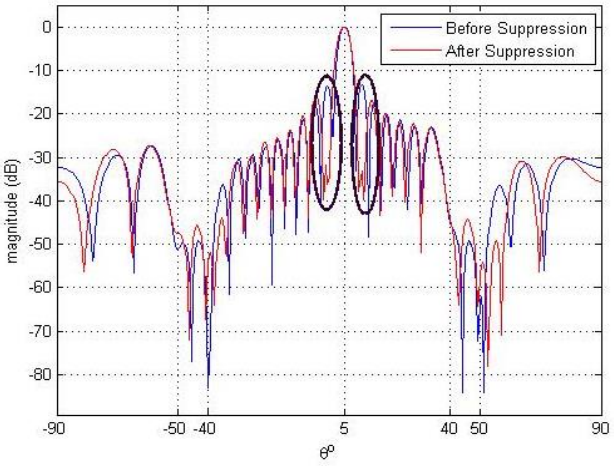


Fig. 6: Beampatterns before and after suppressing the sidelobe levels

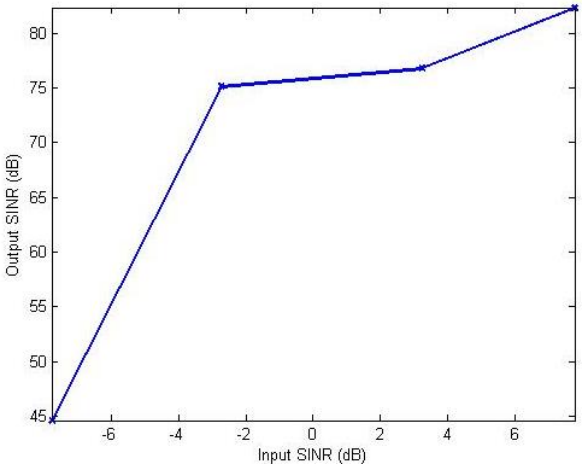


Fig. 7: Input SINR versus output SINR

8. Conclusions

The projected iterative MVDR (PI-MVDR) beamformer is proposed to enhance the conventional MVDR beamformer. The improvement can overcome the inverse covariance matrix avoidance, support available low number of samples, broaden nulls and suppress the first sidelobe levels close to the mainlobe. The covariance matrix used in the PI-MVDR beamformer is an average of between the estimated interference-plus-noise covariance matrix and the sample covariance matrix of the received signal. In the presence of DOA mismatch, low number of samples circumstance and high input SNR, the PI-MVDR is capable of achieving all purposes after a few iterations. In addition, the averaged covariance matrix gives the better results than only using the estimated interference-plus-noise covariance matrix.

9. References

[1] J. R. Guerci, "Theory and application of covariance matrix tapers for robust adaptive beamforming," *IEEE Transactions on Signal Processing*, vol. 47, no. 4, pp. 977-987, Apr. 1999.

[2] W. Li, Y. Zhou, Q. Ye and B. Yang (October 2017). Adaptive antenna null broadening beamforming against array calibration error based on adaptive variable diagonal loading. *International Journal of Antennas and Propagation*. pp. 1-9. Available: http://doi.org/10.1155/2017/3265236

[3] X. Mao, W. Li, Y. Li. Y. Sun and Z. Zhai (2015). Robust adaptive beamforming against signal steering vector mismatch and jammer motion. *International Journal of Antennas and Propagation*. pp. 1-12. Available: http://doi.org/10.1155/2015/780296

[4] I. P. Gravas, Z. D. Zaharias, T. V. Yioultsis, P.I. Lazaridis and T. D. Xenos, "Adaptive beamforming with sidelobe suppression by placing extra radiation pattern nulls," *IEEE Transactions on Antennas and Propagation*, vol. 67, no. 6, pp. 3853-3861, Jun. 2019.

[5] F. Liu, G. Sun, J. Wang and R. Du (April 2014). Null broadening and sidelobe control algorithm via multi-parametric quadratic programming for robust adaptive beamforming. *ACES Journal*, 29(4). pp. 307-315. Available: https://aces-society.org/search.php?vol=29&no=4&type=2

[6] F. Liu, Y. Wu, H. Duan and R. Du (2018). SVR-CMT algorithm for null broadening and sidelobe control. *Progress in Electromagnetics Research*. 163. pp. 39-50. Available: http://www.jpier.org/PIER/pier.php?volume=163

[7] S. A. Vorobyov, "Principles of minimum variance robust adaptive beamforming design," *Signal Processing*, vol. 93, pp. 3264-3277, 2013.

[8] M. Emadi, K.H. Sadeghi, A. Jafargholi and F. Marvasti (2008). Cochannel interference cancellation by the use of iterative digital beam forming method. *Progress in Electromagnetics Research*. 87. pp. 89-103. Available: http://www.jpier.org/PIER/pier.php?volume=87

[9] B. Ors and R. Suleesathira, "First and second order iterative null broadening beamforming," *3rd International*

Conference on Imaging, Signal Processing and Communication, Singapore, Jul. 27-29, 2019.

[10] B. Ors and R. Suleesathira, "Iterative broad null steering," *The 3rd International Conference on Graphics and Signal Processing*, Hong Kong, Jun. 1-3. 2019.

[11] Q. Luo, J. Xie, H. Li and Z. He, "Robust adaptive beamforming in the presence of strong desired signal and DOA mismatch," *International Conference on Computational Problem-solving*, China, Oct. 26-28, 2013.

[12] K. Yang, Z. Zhao and Q. H. Liu (2013). Robust adaptive beamforming against array calibration errors. *Progress in Electromagnetics Research*. 140. pp. 341-351. Available: http://www.jpier.org/PIER/pier.php?volume=140

[13] Y. Hou, L. Xue and Y. Jin, "Robust adaptive beamforming method based on interference-plus-noise covariance matrix," *International Conference on Signal Processing, Communications and Computing*, China, Aug. 5-8, 2013.

ISBN 978-981-14-4787-7
Proceedings of 2020 the 10th International Workshop on Computer Science and Engineering (WCSE 2020)
Yangon (Rangoon), Myanmar (Burma), February 26- February 28, 2020, pp. 103-107

Bilateral Tele-Rehabilitation System with Electrical Stimulation by Using Cloud Service

Yasunori Kawai [1+], Koudai Houga [1], Hiroyuki Kawai [2], and Takanori Miyoshi [3]

[1] Department of Electrical Engineering, National Institute of Technology, Ishikawa College, Japan
[2] Department of Robotics, Kanazawa Institute of Technology, Japan
[3] Department of System Safety, Nagaoka University of Technology, Japan

Abstract. This paper considers a bilateral tele-rehabilitation system with electrical stimulation for a human lower limb by using the Amazon Web Services Internet of Things (AWS IoT) as a cloud service. The scattering matrix method can guarantee the stability for the time delay in the bilateral tele-rehabilitation. This paper applies the AWS IoT to the bilateral tele-rehabilitation system. Because the cross certification and the cipher are provided in AWS IoT, the cybersecurity is considered compared to the HTTP method in the previous research. The experimental results show that the stability can be compensated even if the time delay exist. However, the error between the paddle angle and the knee angle exist.

Keywords: Bilateral Control, Tele-Rehabilitation, Cloud Service

1. Introduction

The electrical stimulation is known as Functional Electrical Stimulation (FES) to improve the motor function of a human. FES makes muscle contraction by external electrical impulses in the same way as electrical impulses from the brain. When the stimulation is controlled, a desired movement can be achieved [1].

The tele-rehabilitation with the electrical stimulation has been proposed in [2]. The tele-rehabilitation indicates that a physical therapist rehabilitates a patient from the remote location. In [2], the therapist uses the haptic device, the lower limb of the patient is controlled by the electrical stimulation. The contact force and position information are communicated between the patient and the therapist. The lower limb of the patient and haptic device are controlled bilaterally. It is called as the bilateral tele-rehabilitation system.

In the previous research, the tele-rehabilitation system with motor-assisted device has been developed in [3]. The physical therapist rehabilitates the patient by the motor-assisted device while the therapist watches the patient through the monitor. It is difficult to feel the condition of the patient. As the another bilateral teleoperation, the tele-rehabilitation system with the haptic device has been developed in [4]. The therapist rehabilitates the patient by grasping the rehabilitation device and the haptic device, respectively. The force and velocity information are communicated between the patient and the therapist. However, the velocity information is not better than the position information in [2].

The stabilization methods for the time delay have been investigated in the bilateral teleoperation, because the bilateral teleoperation is the closed loop system. The passivity-based approach is proposed in [5]-[7] by using the scattering transformation and the wave variable. The stability is guaranteed by the passivity of the operator, the tele-operator, and the communication block including time delays. As one of the non-passive approach for the time delay, the scattering matrix method has been proposed in [2],[8]. The stability is guaranteed based on the norm of the operator and the tele-operator.

+ Corresponding author. Tel.: +81-76-288-8110
E-mail address: y_kawai@ishikawa-nct.ac.jp

This paper considers the experimental verification of the bilateral tele-rehabilitation system using Amazon Web Services Internet of Things (AWS IoT). The main contribution is to apply the AWS IoT to the bilateral tele-rehabilitation system. Because the cross certification and the cipher are provided in AWS IoT, the cybersecurity is considered compared to the HTTP method in the previous research in [9]. The stability is confirmed when the time delay exists by the AWS IoT in the experiment.

2. Bilateral Tele-rehabilitation System Using Scattering Matrix Method

This section reviews the bilateral tele-rehabilitation system based on the scattering matrix method in the previous work of [2]. Fig. 1 shows the bilateral tele-rehabilitation system of the human lower limb. A force $f_h(t) \in R$ is provided to the paddle by the therapist. The force information is sent to a patient through the Internet. In the patient side, the electrical stimulus signal is made according to the received force information. The stimulus signal is provided to the quadriceps femoris muscle group e_2 and e_3. The muscle contraction induced by the electrical stimulation yields the leg extension. At the same time, the information of the patient's knee angle $q_s(t) \in R$ is sent to the therapist through the Internet. In the therapist side, the paddle's angle $q_m(t) \in R$ is controlled to be the same angle as the patient's knee angle $q_s(t)$ by the motor. Then, the therapist can feel the motion of the patient's lower limb by the reaction force from the paddle.

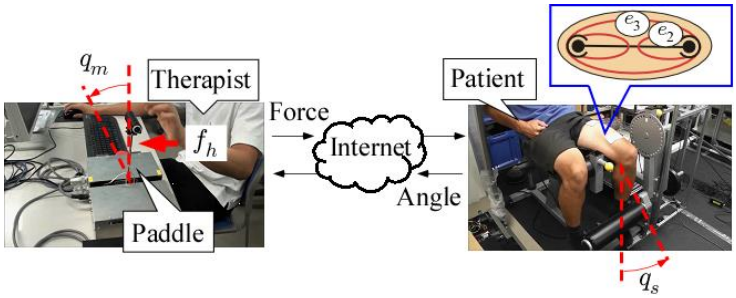


Fig. 1: Bilateral tele-rehabilitation system of the human lower limb.

The bilateral tele-rehabilitation system using the scattering matrix method is illustrated in Fig. 2. The therapist model is $G_{mm}(s)$, the patient model is $G_{ss}(s)$, the wave filters are $W_m(s)$ and $W_s(s)$, time delays are $T_1(t)$ and $T_2(t)$ in the communication by the Internet. The force $f_h(t)$ is indicated in Fig. 1, the disturbance is $f_e(t) \in R$.

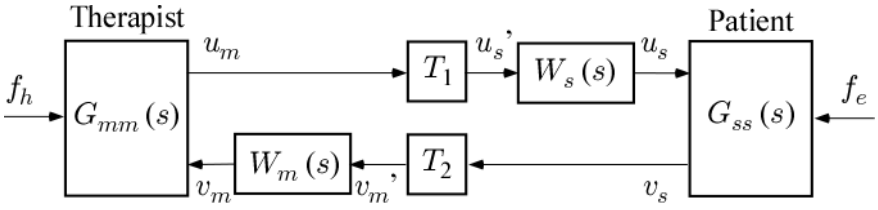


Fig. 2: Bilateral tele-rehabilitation system using the scattering matrix method.

The therapist model $G_{mm}(s)$ in Fig. 2 is shown in Fig. 3. The paddle model is $P_m(s)$, the friction force model of the paddle is $P_f(s)$, the feedback controller is $K_m(s)$, the phase-lag filter is $G_{cm}(s)$, the square box which is drawn by the dashed line is the scattering transformation, where $b \in R$ is the design parameter. The paddle angle is $q_m(t)$, the reference angle of the paddle angle is $q_{rm}(t) \in R$ which is equal to the past patient's knee angle $q_s(t - T_2)$, the force which is measured by the force sensor on the paddle is $f_m(t) \in R$.

The patient model $G_{ss}(s)$ in Fig. 2 is shown in Fig. 4. The patient's lower limb model is $P_s(s)$, the phase-lead filter is $G_{cs}(s)$, the square box written by the dashed line is the scattering transformation. The phase-lead filter $G_{cs}(s)$ is needed to stabilize the patient, because the feedback controller doesn't exist in the patient side. The patient's knee angle is $q_s(t)$, the force $f_s(t) \in R$ is equal to the force $f_m(t - T_1)$. The force $f_s(t)$ is provided by the muscles e_2 and e_3 by the electrical stimulation.

The stability condition of the bilateral teleoperation system based on the scattering matrix method has already been proposed in [8] as follows by using the small gain theorem

$$\|G_{mm}(s)W_m(s)\|_\infty \cdot \|G_{ss}(s)W_s(s)\|_\infty \leq 1. \tag{1}$$

The stability condition Eq. (1) means that the stability condition can be evaluated by using H_∞ norm regardless of the length of the time delay. The control objectives are indicated as follows;

(1) The paddle's angle $q_m(t)$ is the same angle as the patient's knee angle $q_s(t)$.

(2) The therapist feels the reaction force from the paddle.

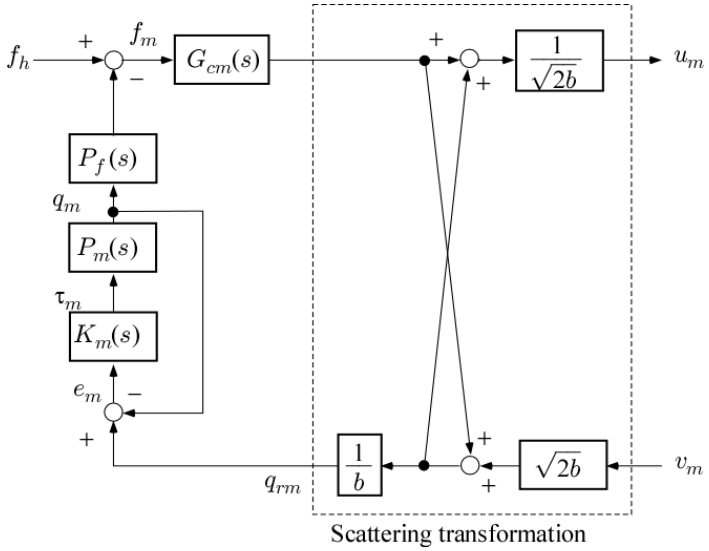


Fig. 3: Therapist model $G_{mm}(s)$ in Fig. 2.

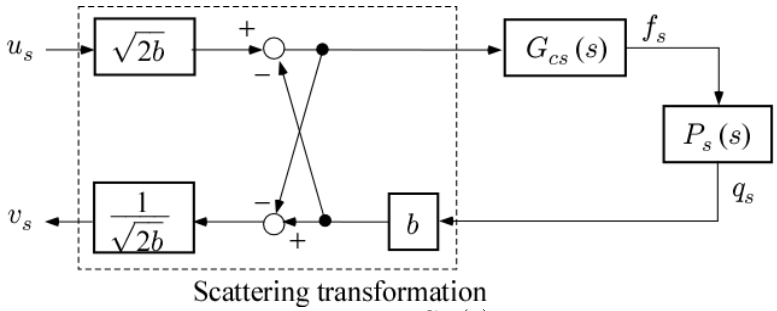


Fig. 4: Patient model $G_{ss}(s)$ in Fig. 2.

3. Experimental Setup

The experimental equipment of the tele-rehabilitation system is shown in Fig. 5. In the therapist side, the contact force $f_h(t)$ is measured by using the force sensor where the force sensor value is $f_m(t)$. The paddle angle $q_m(t)$ and reaction force $\tau_m(t)$ are obtained by utilizing the encoder and DC motor, respectively. In the patient side, the patient sits on the leg extension machine with a weight 5 [LBS] (=2.3 [kg]). The knee angle $q_s(t)$ is measured by using the encoder. The contact force $f_s(t)$ is converted to the electrical stimulus signal $\tau_s(t)$ by utilizing the RehaStim (HASOMED GmbH). The PC with the signal processing board Q8-USB (Quanser) is connected to the therapist side and patient side. The communication between the PC and the AWS IoT are implemented by the "Publish" and "Subscribe". In the AWS IoT, the things Topic1 and Topic2 are made. The information $f_m(t)$ and $q_s(t)$ are sent to the Topic 1 and Topic2 by "Publish". When the "Publish" is implemented, the PC takes the information $f_s(t)$ and $q_s(t)$ by "Subscribe". The design parameters are designed as follows;

$$G_{cm}(s) = \frac{0.1s+5}{s+5},\ G_{cs}(s) = G_{cm}^{-1}(s) = \frac{s+5}{0.1s+5}, \tag{2}$$

$$K_m(s) = 2 + \frac{2}{s} + 0.1s, \tag{3}$$

$$W_m(s) = W_s(s) = \frac{1}{(0.05s+1)},\quad b = 20. \tag{4}$$

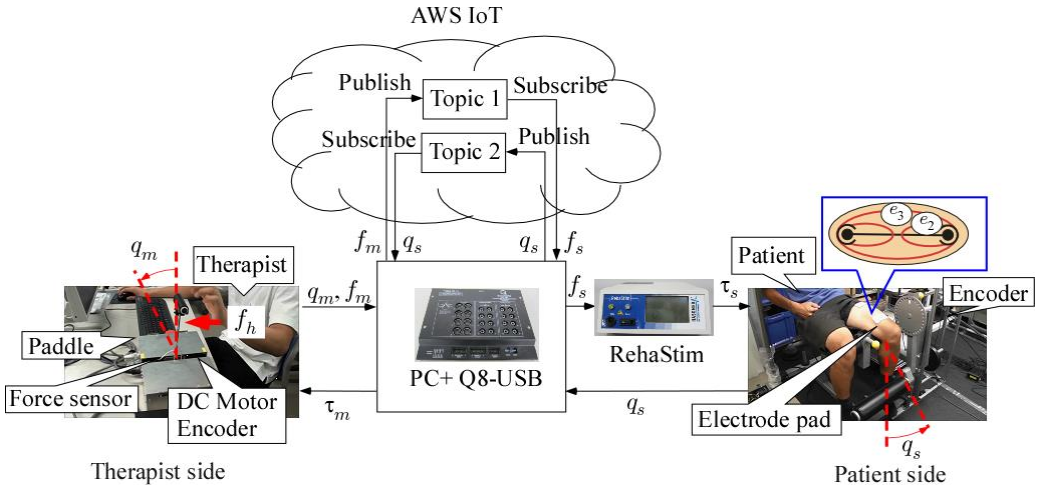


Fig. 5: Experimental equipment.

4. Experimental Result

The experimental results are shown in Figs. 6-10. The therapist provides the contact force $f_h(t)$ to the paddle. The sensor value $f_m(t)$ is indicated in Fig. 6, where the dashed line is $f_m(t)$. The solid line is $f_s(t)$ which is same as $f_m(t-T_1(t))$ in Fig. 6. The forces $f_s(t)$ and $f_m(t)$ almost have the same value. The magnitude of the electrical stimulation $\tau_s(t)$ according to the force $f_s(t)$ is illustrated in Fig. 7. Then, the knee angle of the patient $q_s(t)$ is illustrated in Fig. 8, where the solid line is $q_s(t)$. The dashed line in Fig. 8 is the paddle angle $q_m(t)$. The reaction force $\tau_m(t)$ is shown in Fig. 9. The time delays are indicated in Fig. 10. The time delays $T_1(t)$ and $T_2(t)$ are the required time from the therapist to the patient and the patient to the therapist, respectively. The average time delay is 0.2 [s]. Though the error between the paddle angle $q_m(t)$ and the knee angle $q_s(t)$ exists for $3 \leq t \leq 5$, the stability is compensated even if the time delay exist.

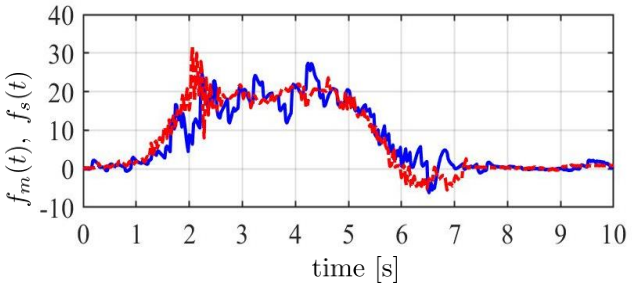


Fig. 6: Time response of the force $f_m(t)$ and the force $f_s(t)$. (solid: $f_s(t)$, dashed: $f_m(t)$)

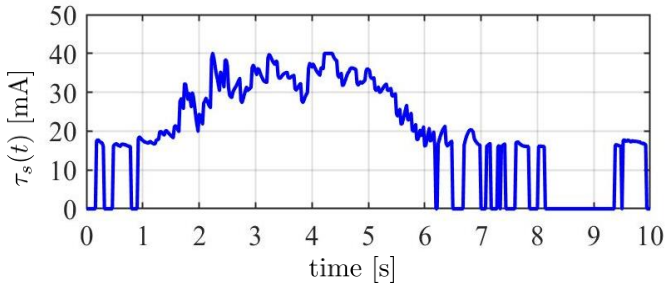


Fig. 7: Time response of the magnitude of the electrical stimulation $\tau_s(t)$.

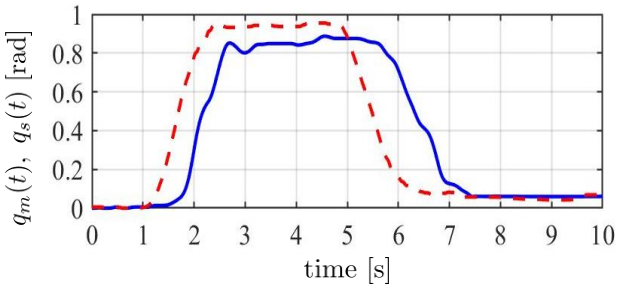


Fig. 8: Time response of the paddle angle $q_m(t)$ and the knee angle $q_s(t)$. (solid: $q_s(t)$, dashed: $q_m(t)$)

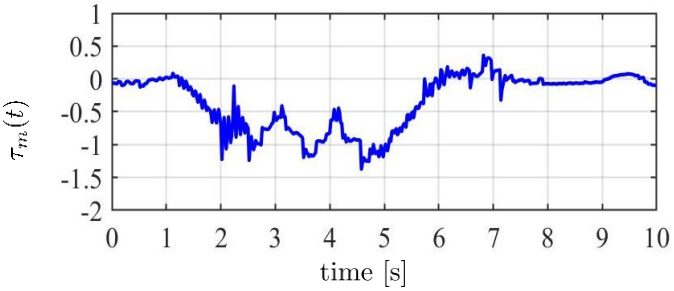


Fig. 9: Time response of the reaction force $\tau_m(t)$.

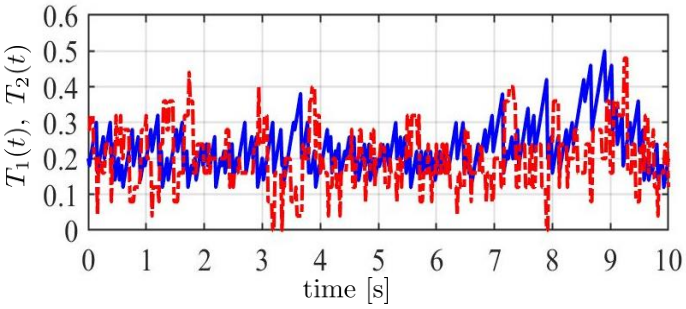


Fig. 10: Time response of the time delays $T_1(t),\ T_2(t)$. (solid: $T_1(t)$, dashed: $T_2(t)$)

5. Conclusion

This paper considered the experimental verification of the bilateral tele-rehabilitation system using AWS IoT. The therapist side and the patient side can be communicated through two things on AWS IoT. The error between the paddle angle and the knee angle exist. However, the stability can be compensated for the average time delay 0.2 [s].

6. References

[1] C. T. Freeman, E. Rogers, A. Hughes, J. H. Burridge, and K. L. Meadmore, "Iterative Learning Control in Health Care: Electrical Stimulation and Robotic-Assisted Upper-Limb Stroke Rehabilitation," *IEEE Control Systems Magazine*, Vol. 32, No. 1, pp. 18-43, 2012.

[2] Y. Kawai, K. Honda, H. Kawai, T. Miyoshi, and M. Fujita, "Tele-Rehabilitation System for Human Lower Limb using Electrical Stimulation based on Bilateral Teleoperation," in *Proc. the 2017 IEEE Conference on Control Technology and Applications (CCTA)*, pp. 1446-1451, 2017.

[3] J. Bae, W. Zhang, and M. Tomizuka, "Network-Based Rehabilitation System for Improved Mobility and Tele-Rehabilitation," *IEEE Trans. on Control Systems Technology*, Vol. 21, No. 5, pp. 1980-1987, 2013.

[4] S. F. Atashzar and M. Shahbazi and M. Tavakoli, and R. V. Patel, "A Passivity-Based Approach for Stable Patient-Robot Interaction in Haptics-Enabled Rehabilitation Systems: Modulated Time-Domain Passivity Control," *IEEE Trans. on Control Systems Technology*, Vol. 25, No. 3, pp. 991-1006, 2017.

[5] T. Hatanaka and N. Chopra and M. Fujita, and M. W. Spong, *Passivity Based Control and Estimation in Networked Robotics*, Springer International Publishing, Communications and Control Engineering Series, 2015.

[6] P. F. Hokayem and M. W. Spong, "Bilateral Teleoperation: An Historical Survey," *Automatica*, Vol. 42, No. 12, pp. 2035-2057, 2006.

[7] E. Nuno and L. Basanez, and R. Ortega, "Passivity-based control for bilateral teleoperation: A tutorial," *Automatica*, Vol. 47, No. 3, pp. 485-495, 2011.

[8] T. Miyoshi, K. Terashima, and M. Buss, "A Design Method of Wave Filter for Stabilizing Non-Passive Operating System," in *Proc. the 2006 IEEE International Conference on Control Applications*, pp. 1318-1324, 2006.

[9] Y. Kawai, K. Honda, and M. Koshino, "Bilateral Tele-Rehabilitation System with Electrical Stimulation through Cloud Server," in *Proc. the 2019 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW)*, 2019.

ISBN 978-981-14-4787-7
Proceedings of 2020 the 10th International Workshop on Computer Science and Engineering (WCSE 2020)
Yangon (Rangoon), Myanmar (Burma), February 26- February 28, 2020, pp. 108-113

Influences of Network Delay and Moving Velocity on Virtual Cooperative Work with Haptic Sense

May Zin Oo[1+], Yutaka Ishibashi[2], Khin Than Mya[3]

[1, 3]University of Computer Studies, Yangon, Myanmar
[2] Nagoya Institute of Technology, Nagoya, Japan

Abstract. In this paper, we investigate influences of the network delay and moving velocity on virtual cooperative work with haptic sense by experiment. In the work, each of two users operates a haptic interface device, and the two users collaboratively raise a stick in a 3D virtual space. The stick has a weighted ball, and the ball moves to the lower end along the stick if one end of the stick is lower than the other end. They try to keep the ball at the center of the stick while raising the stick. Experimental results demonstrate that it is possible to keep the ball around at the center when the network delay is small. We also show that there exists the optimal moving velocity of the ball depending on the network delay.

Keywords: virtual cooperative work, haptic sense, network delay, moving velocity, experiment.

1. Introduction

A number of researchers have been studying networked virtual environments with haptic sense [1]-[4]. We can largely improve the efficiency of collaborative work by using haptic sense. In such work, it is necessary for multiple users to do collaborative work while watching the same displayed images in a 3D virtual space simultaneously. However, when the information about the space is transmitted over a Quality of Service (QoS) [5] non-guaranteed network like the Internet, the receiving times of the information at different terminals may be different from each other owing to network delays, delay jitters, and so on. That is, some of the terminals may have already received information while the others may have not received the information yet. Then, the efficiency of collaborative work may deteriorate seriously because users at the terminals may watch different displayed images at the same time.

To solve the above problem, as QoS control, we need to perform simultaneous output-timing control [6], which absorbs delay differences among different terminals. Various types of the control such as the local lag control [7], the group synchronization control [8], and the adaptive Δ-causality control [9] can be employed.

In [2], Arimoto et al. deal with a networked music ensemble system with haptic sense. A pair of users play drums. In the system, the network delay and packet loss degrade collaboration performance between the users. They examine the influence of network delay, the effect of media synchronization control, and the role of haptic sense in the system. As a result, they illustrate that using haptic sense in addition to the visual and auditory senses can enhance the collaboration performance and improve the sense of togetherness among individuals taking part in the work. However, they investigate only the influence of network delay in the work; they do not examine influences of other factors such as the velocity of hitting drums.

In [4], the authors handle a networked balance system with haptic sense in which each of two users operates a haptic interface device, and the two users collaboratively raise a stick by lifting two ends of the stick up in a 3D virtual space. The stick has a weighted ball, and the ball moves to the lower end along the stick if one end of the stick is lower than the other end. They try to keep the ball at the center of the stick

+ Corresponding author. Tel.: +81-76-288-8110
E-mail address: mayzinoo2018@ucsy.edu.mm

while raising the stick. They carry out the local lag control as simultaneous output-timing control, and they make a comparison between results with the local lag control and those without the control to clarify the effect of the control on the human perception of weight. However, the influences of the other factors such as the moving velocity of the ball on human perception of weight have not been investigated. We need to examine influences of the moving velocity as well as the network delay on the collaborative work to carry out QoS control more efficiently.

In this paper, we deal with the networked balance system with haptic sense [4], and we investigate the influences of the network delay and moving velocity on the work efficiency. We also examine the influence of the initial position of the ball.

The remainder of this paper is organized as follows. Section 2 outlines the networked balance system with haptic sense, and Section 3 describes the experiment method. Then, Section 4 presents experimental results and discusses them. Finally, Section 5 concludes the paper.

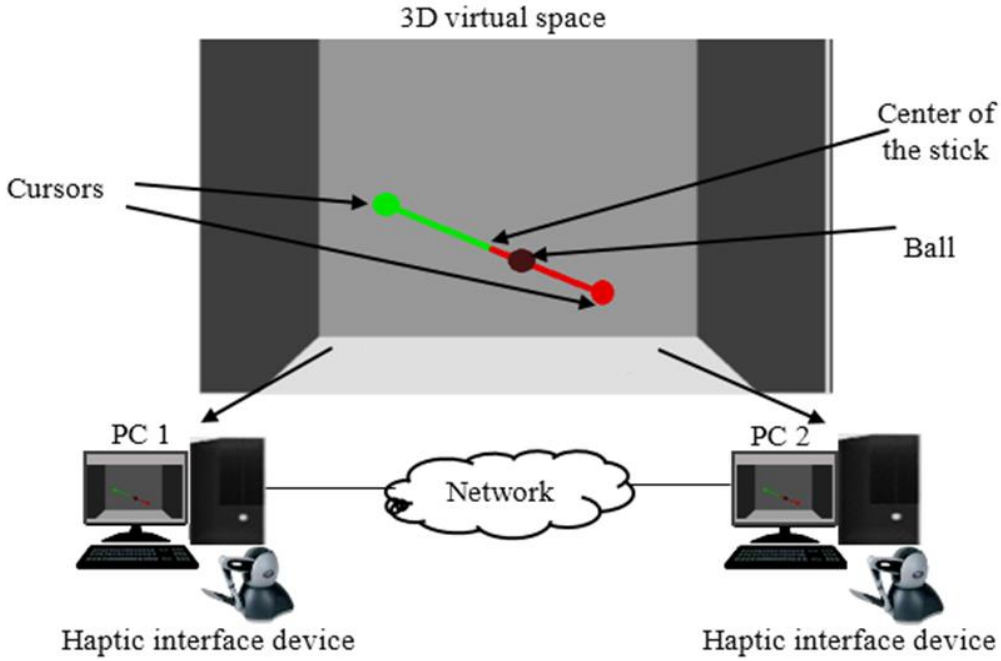


Fig. 1: Configuration of networked balance system with haptic sense.

2. Networked Balance System with Haptic Sense

We show the configuration of the networked balance system with haptic sense and a displayed image of the 3D virtual space in Fig. 1. The system consists of two terminals (called *PCs 1 and 2* here) each of which has a haptic interface device (3D Systems Touch [10]).

By operating the haptic interface device, a user of each PC can move a cursor which denotes the position of the device stylus's tip in the space which is surrounded by a floor, a ceiling, and walls. The two cursors are fixed at two ends of a stick whose weight is 0 gf. The stick can be stretched or shrunk freely between the two ends. A ball whose weight is 270 gf is placed on the stick between the two cursors (i.e., the two ends of the stick). The reaction force applied to a user is proportional to the distance from the user's cursor to the ball (as shown in the 3D virtual space of Fig. 1, the right side of the center is green, and the left side of the center is red). We show the calculation method of reaction force in Fig. 2, where the proportion of the distance from the red cursor to the ball to the distance from the green cursor to the ball is set to k : $(1.0 - k)$. Also, m is the mass of the ball, and $\boldsymbol{g}$ (9.8 m/s^2) is the gravitational acceleration. The reaction force applied to the red cursor's user is $(1.0 - k)\ \boldsymbol{F}$, and that applied to the green cursor's user is $k\ \boldsymbol{F}$. The ball moves towards the lower cursor along the stick if there exist altitude differences between the two cursors. In this paper, we assume that there is no viscous resistance. The users try to keep the ball at the center of the stick while lifting up the stick.

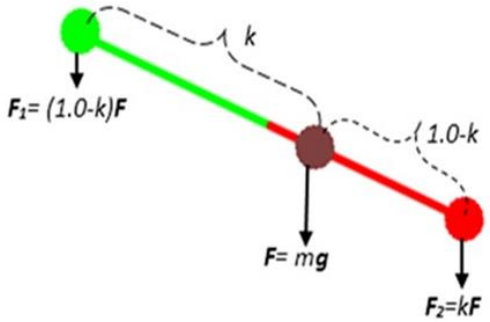


Fig. 2: Calculation method of reaction force.

3. Experiment Method

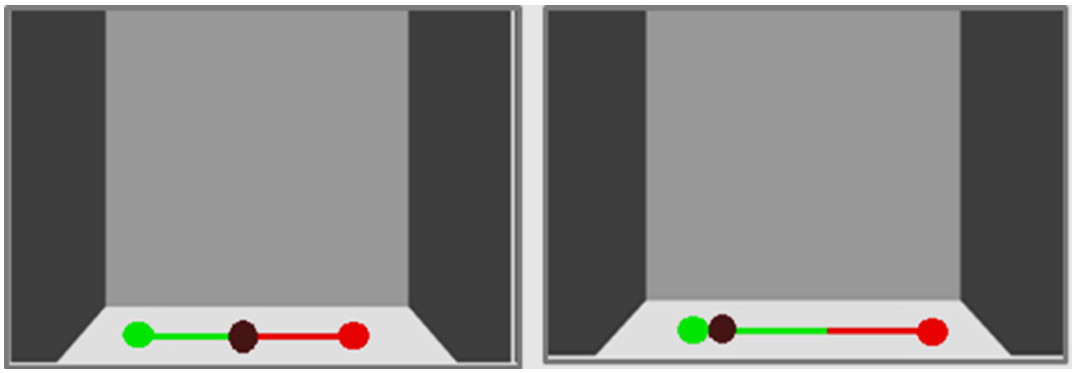
In our experiment system, we employed a network emulator (netem [11]) instead of the network in Fig. 1 to generate a constant delay (called the *additional delay* in this paper) for each packet transmitted between the two PCs. Note that we here handle constant delay for simplicity because media synchronization control absorbs the fluctuation of the network delay to some extent. In the experiment, we handled two methods (called *Methods 1 and 2*). In Method 1, the users started to lift up the ball from the initial state in which the weighted ball was placed at the center of the stick, and the two cursors were placed on the floor (see Fig. 3 (a)). The initial position of the ball in Method 2 is different from that in Method 1; the ball was placed at the far left of the stick (see Fig. 3 (b)). In both methods, the users were asked to lift the ball up to a height of 16.7 (we assume that the diameter of cursor is 1 in this paper) collaboratively while trying to keep the ball at the center of the stick and to move at a constant velocity (see Fig. 4). The movement distance of 16.7 corresponds to 10 cm in the real space when each user raises the stylus of the haptic interface device vertically.

If the ball was not located at the center, each user tried to adjust the place of the ball by moving up his/her cursor or stop moving. Then, we investigated the influence of the additional delay and moving velocity, and we also examine whether the optimal velocity exists for each additional delay.

We carried out the experiment in which the additional delay from PC 1 to PC 2 was set to the same value as that from PC 2 to PC 1, and the additional delay was changed from 0 ms to 150 ms at intervals of 50 ms. Also, we changed the moving velocity from 0.1 to 0.9 at intervals of 0.2 and from 1.0 to 13.0 at intervals of 2.0.

The positions of the ball and the operation time were measured in the experiment. We define the position of the ball as the distance between the ball and the center of the stick. Positions on the right side of the center are denoted by plus values, and those on the left side are done by minus ones. The operation time is defined as a time interval from the moment one user starts to raise the cursor until the instant one user's cursor reaches the maximum height of 16.7.

In the experiment, in both methods, the combinations of the additional delay and moving velocity were selected in random order for each work. The experiment was conducted by two users (females) whose ages were 35. We carried out the experiment 10 times.



(a) Ball at center (Method 1) (b) Ball at far left (Method 2)

Fig. 3: Displayed images of initial state in experiment.

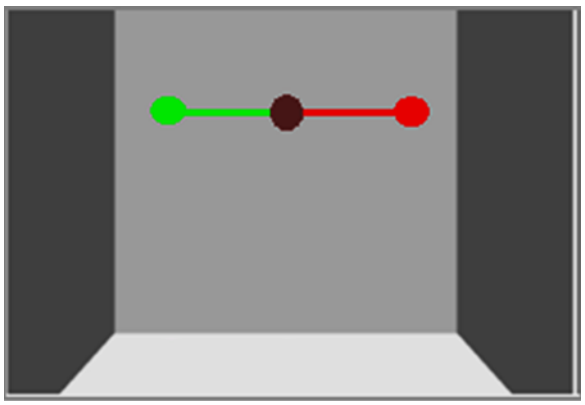


Fig. 4: Displayed image of end state in experiment.

4. Experimental Results

We show the average of average absolute position (i.e., the mean of 10-times average absolute position) versus the moving velocity at PC 1 and PC 2 in Figs. 5 (a) and (b), respectively, for Method 1. Figure 6 shows the average operation time versus the additional delay for Method 1. Figures 7 (a) and (b) plot the average of average absolute position versus the moving velocity at PCs 1 and 2, respectively, for Method 2. Figure 8 shows the average operation time versus the additional delay for Method 2.

In Fig. 5, we see that in Method 1, as the additional delay and moving velocity increase, the average of average absolute position becomes larger. This means that it is difficult to keep the ball at the center of the stick when the moving velocity is faster than around 10. We also notice in the figure that the average of average absolute position is smaller than about 2 for the velocity less than around 4; in this case, the ball was almost kept at the center. We further observe that the results of PC 1 are somewhat different from those of PC 2. We can adjust the positions at PCs 1 and 2 by performing simultaneous output-timing control; this is our further study.

From Fig. 6, we find that the average operation time increases as the moving velocity and additional delay become larger. Therefore, the users tried to adjust the ball many times when the moving velocity and additional delay were large.

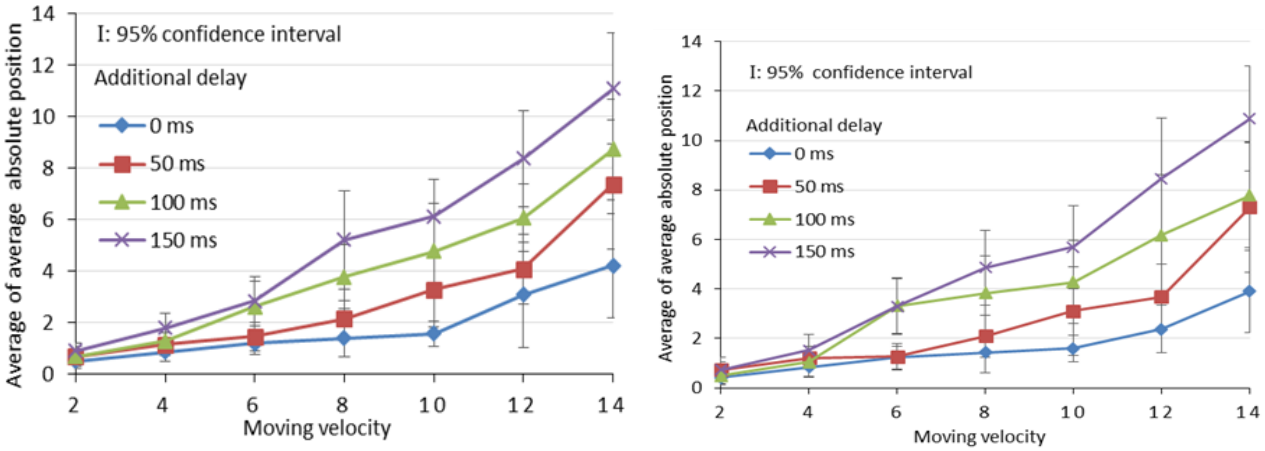
From Fig. 7, we notice that there exists the optimal moving velocity for each additional delay in Method 2. It should be noted that the optimal moving velocity has the smallest average of average absolute position. Thus, the users can more easily adjust the ball at the center of the stick at the optimal velocity. This is because when the moving velocity is too small, it takes a long time to move the ball to the center, and when the moving velocity is too large, it is difficult to keep the ball at the center of the stick. It should be noted that smaller values of the moving velocity are better in Method 1. Furthermore, we confirm in Fig. 7 that the results of PC 1 are largely different from those of PC 2 when the moving velocity is larger than about 3.0.

Figure 8 reveals that the average operation time becomes shorter as the moving velocity increases. This is because we can easily keep the ball at the center by lifting the cursors up timely when the initial position of the ball is far left; that is, the user at PC 1 lifts the cursor at first, and then the user at PC 2 lifts the cursor while watching the ball so that the ball can stops at the center. Therefore, the influence of the initial position of the ball on the operation time is large.

5. Conclusions

In this paper, we investigated the influences of the network delay and moving velocity on the work efficiency in the networked balance game where two users collaboratively lift up a weighted ball with haptic sense in a virtual environment. We also examined the influence of the initial position of the ball. As a result, we found that it is possible to keep the ball around at the center when the network delay is small. Also, there exists the optimal moving velocity depending on the network delay.

As our next step, we will examine influences of other factors such as the weight and size of the ball on the work efficiency. It is also important to make a mathematical model of our system and to analyze the results in this paper.



(a) PC 1 (b) PC 2

Fig. 5: Average absolute position versus moving velocity in Method 1.

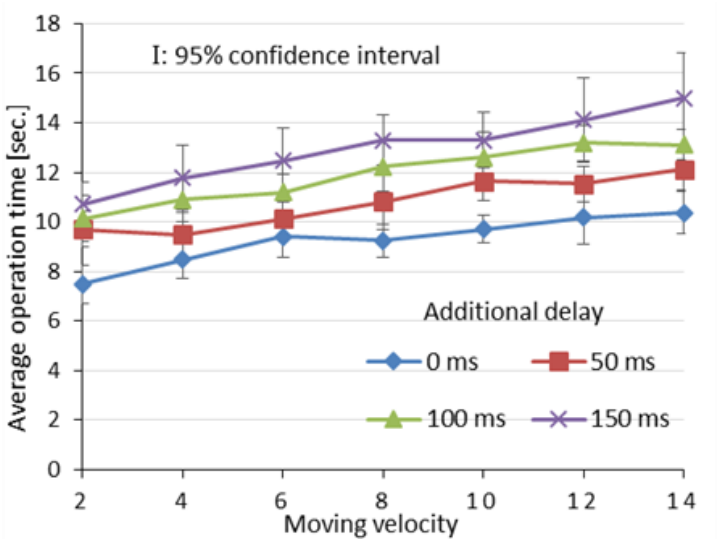
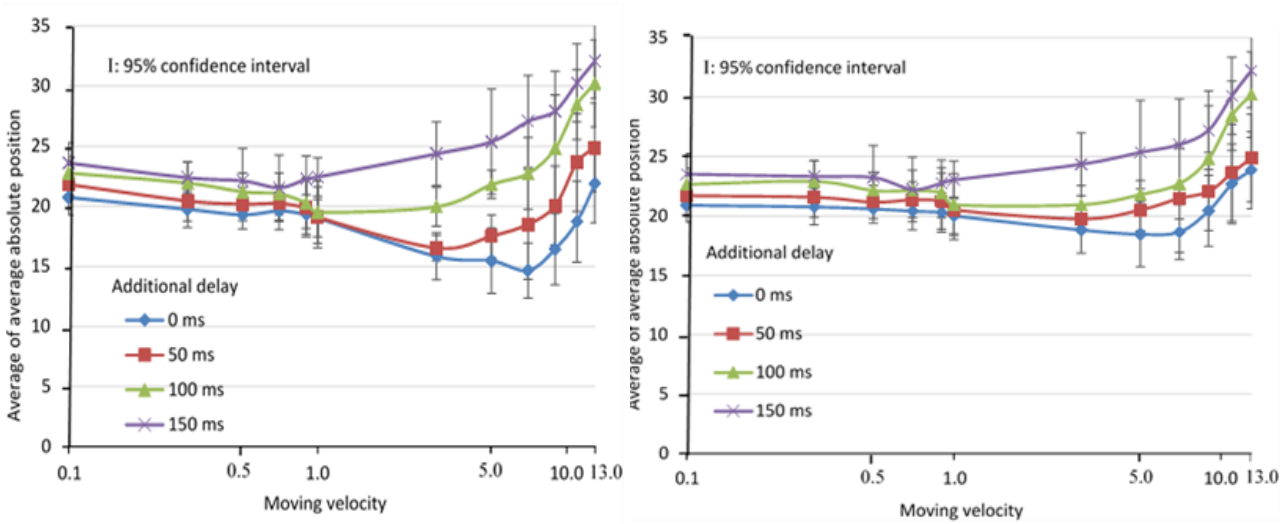


Fig. 6: Average operation time versus moving velocity in Method 1.



(a) PC 1 (b) PC 2

Fig. 7: Average of average absolute position versus moving velocity in Method 2.

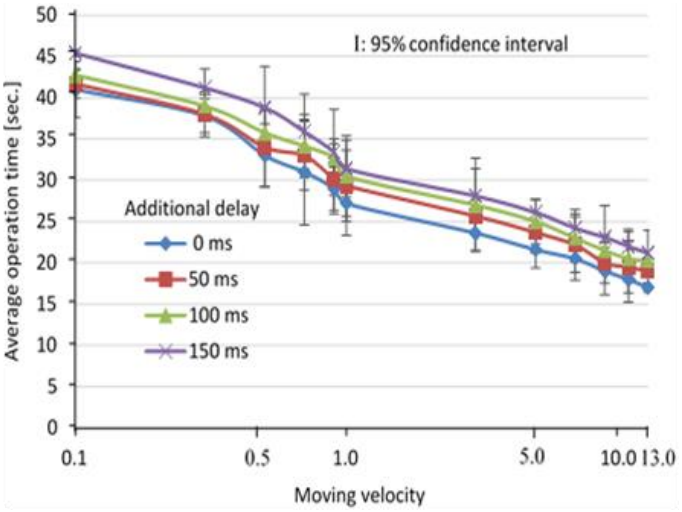


Fig. 8: Average operation time versus moving velocity in Method 2.

6. References

[1] E. Steinbach, S. Hirche, M. Emst, F. Brandi, R. Chaudhari, J. Kammerl, and I. Vittorias, "Haptic communications," IEEE Journals & Magazines, vol. 100, no. 4, pp. 937-945, Apr. 2012.

[2] I. Arimoto, K. Hikichi, K. Sezaki, and Y. Yasuda, "Influence of network delay on ensemble application" in Proc. IEEE International Workshop on Haptic Audio Visual Environments and their Applications (HAVE), pp. 1-2, Oct. 2005.

[3] B. M. Lambeth, J. LaPlant, E. Clapan, and F. G. Hamza-Lup, "The effects of network delay on task performance in a visual-haptic collaborative environment," in Proc. ACM The 47th Annual Southeast Regional Conference, pp. 1-5, Mar. 2009.

[4] P. Huang and Y. Ishibashi, "Human perception of weight in networked virtual environment with haptic sense: Influence of network delay," Journal of Communications (JCM), vol. 14, no. 6, pp. 478-483, June 2019.

[5] ITU-T Rec. I. 350 "General aspects of quality of service and network performance in digital networks," Mar. 1993.

[6] P. Huang, Y. Ishibashi, and M. Sithu, "Enhancement of simultaneous output-timing control with human perception of synchronization errors among multiple destinations," in Proc. The 2nd IEEE International Conference on Computer and Communications (ICCC), pp. 2099-2103, Oct. 2016.

[7] M. Mauve, J. Vogel, and W. Effelsberg, "Local lag and timewarp: Providing consistency for replicated continuous applications," IEEE Trans. on Multimedia, vol. 6, no. 1, pp. 47-57, Feb. 2004.

[8] Y. Ishibashi, A. Tsuji, and S. Tasaka, "A group synchronization mechanism for stored media in multicast communications," in Proc. IEEE International Conference on Computer Communications (ICCC), pp. 693-701, Apr. 1997.

[9] Y. Ishibashi, S. Tasaka, and Y. Tachibana, "Adaptive causality and media synchronization control for networked multimedia applications," in Conf. Rec. IEEE International Conference on Communications (ICCC), pp. 952-958, June 2001.

[10] [Online] Available: https://www.3dsystems.com/haptics-devices/touch.

[11] [Online] Available: http://www.linux.org/man8/tc-netem.

ISBN 978-981-14-4787-7
Proceedings of 2020 the 10th International Workshop on Computer Science and Engineering (WCSE 2020)
Yangon (Rangoon), Myanmar (Burma), February 26- February 28, 2020, pp. 114-121

An Investigation on Stability and Operability in Haptic Communication Systems

Hitoshi Watanabe [1+], Pingguo Huang [2] and Yutaka Ishibashi [3]

[1] Faculty of Engineering, Tokyo University of Science, Tokyo, Japan
[2] Faculty of Business Management, Seijoh University, Tokai, Japan
[3] Graduate School of Engineering, Nagoya Institute of Technology, Nagoya, Japan

Abstract. In the remote robot systems using haptic communication under communication delays, there is a problem of instability of operations or behaviors such as mechanical howling. For solving the problem, we have introduced a method using "resistance" which is proportion to the velocity and investigated the theoretical method to analyze the mechanism by using difference differential equations. In our previous theoretical research, we have clarified that the resistance has the effect to decrease the mechanical howling but it degrades the operability of the action to approach to the target position. These results suggest that the minimum resistance which does not cause the mechanical howling is the optimal resistance value which keeps the operability of the target approaching action high as much as possible. This study investigates the feasibility of this method by experiment and theoretical analysis.

Keywords: Haptic communication; Communication delay; Stability; Operability; Resistance; Difference differential equation; Human characteristics

1. Introduction

The haptic communication attracts great attention now [1] and applications for many fields, e.g. tele-surgery, are being investigated [2]. However, when a remote robot system with haptics is executed through a network in which the QoS (Quality of Service) [3] is not guaranteed like as the Internet, there is the possibility that the QoE (Quality of Experience) [4] may degrade heavily. For solving this problem, we have introduced a method using "resistance" [5] which is proportion to the velocity and investigated the effects of this method by experiment using real systems. Moreover, we have investigated the theoretical analysis [6], [7], [8] to clarify the mechanism by using difference differential equation [9], [10]. In our previous theoretical research, we have clarified that the resistance has the effect to decrease the mechanical howling but it degrades the operability of the action to approach the target position [11]. Moreover, we have also investigated the basic frame to determine the optimal resistance based on the estimation of human operation characteristics by experiment [8]. This paper investigates the feasibility of the method to keep the operability high as much as possible and suppress the mechanical howling by choosing the optimal value of the resistance.

The new theoretical consideration is added to the previous research and experimental trial to clarify the feasibility of the proposed method has been done by using the haptic communication system.

2. The Behavior of Remote Robot Systems

Fig. 1 shows the behaviour of the remote robot systems. The moving of the object is executed by determining the position at the next time by the master's instruction. The reaction force which is required for

[+] Corresponding author. Tel.: +81-3-5876-1381
E-mail address: kwata@rs.tus.ac.jp

moving the object under the certain kinetic law is returned to the master as the "reaction force". Here, S_t is the position of slave at time t, M_t is the instructed position by the maser at time t and F_t is the reaction force.

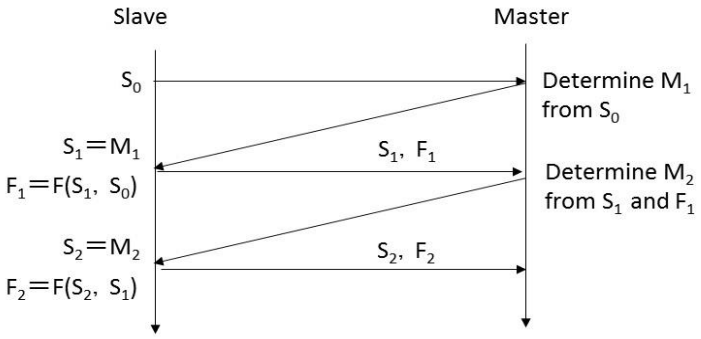


Fig.1: Behaviour of Remote Robot System

The resistance which we introduced is the following coefficient C_d converting the M_t to the S_t.

$$S_{t+1} = M_t - Cd(Mt - St)$$
$$0 \leq Cd < 1 \tag{1}$$

The operator manipulates as to approach the object to the target position. This action is done by considering the present position of the object and the reaction force. This is the feedback loop with time delay. On the other hand, for the feedback of reaction force from slave to master, it is necessary to transform the reaction force to some displacement at the master. One of its methods is to use the spring-dumper mechanism. This is another feedback loop with time delay.

3. The Feedback Loop for Target Approaching Action

Let $y(t)$ be the position of the object at time t, V be the velocity of manipulation for approaching to the target position and, y_D be the target position. Also V is assumed as a function of $y(t)$-y_D as formula (2) [6].

$$V(t) = V(y(t) - y_D) \tag{2}$$

V is the velocity of the manipulating position but not match to the velocity of the position of the object, i.e., in the case where the resistance C_d is introduced. Furthermore, in the case where the communication delay is δ, V is determined based on the position of object time δ ago. Therefore, the following formula is obtained.

$$\frac{dy(t)}{dt} = F(V(y(t-\delta) - y_D)) \tag{3}$$

Here, F is the function which associates the position of the master and one of the slaves. Formula (3) is a difference differential equation. Assuming that the position of the master matches to the one of the slaves and the velocity function has formula (4), the behaviour of the solution depends α and δ. An example is shown in Fig 2.

$$V(y) = A(y - yd)^{\alpha} \tag{4}$$

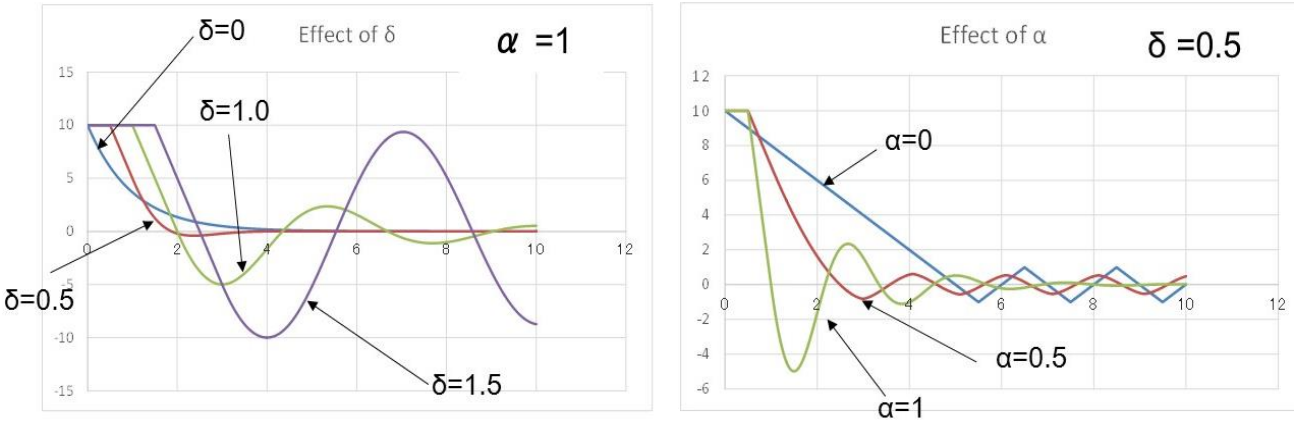
Whether the solution vibrates or oscillates or not is the important point to evaluate the stability and operability. It is decided by the examination of the characteristic function, if V is a linear function. Let $y(t) = C\, e^{-\lambda t}$ be a solution of the following formula (5).

$$\frac{dy(t)}{dt} = -Ay(t-\delta) \tag{5}$$

λ should satisfy formula (6).

$$\lambda\delta + A\delta e^{-\lambda\delta} = 0$$
$$\therefore \quad u + ke^{-u} = 0 \tag{6}$$

Here, k= δ /*A*. Formula (6) is called the characteristic equation. Examining the positive or negative of the real part of the roots of formula (6), the stability of the solution of formula (5) can be decided [10]. For example, the solution does not vibrate and converge to 0 when $\delta < (1/e)(1/A)$, and the solution vibrates and diverges to infinity when $\delta > \pi /A$ (see Fig.3).



(a) Effect of communication Delay (b) Effect of function V

Fig. 2: Effect of communication delay and velocity function

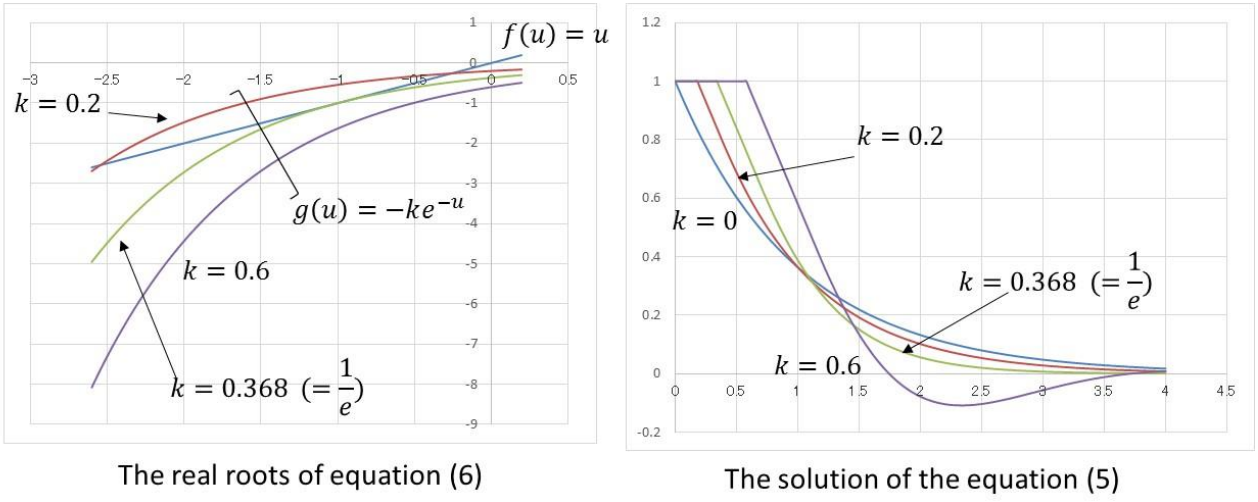


Fig. 3: The behaviour of the solution of (5)

We have applied this method for evaluating the stability when the resistance C_d is introduced. When the control period of the remote robot system is adequately small, the input function *g(t)* is converted to the output function *f(t)* by formula (7) [7]. It means that C_d is a kind of the first-order delay element. Here, $k= \Delta (1-C_d)/C_d$ and Δ is the sampling period. The difference differential equation is obtained as formula (8). The characteristic equation is obtained as formula (9).

$$f(t) = ke^{-kt} \int_{-\infty}^{t} e^{ku} g(u) du \tag{7}$$

$$\tau \frac{d2f(t)}{dt2} + \frac{df(t)}{dt} + Af(t-\delta) = 0$$

$$\text{here } \tau = \frac{1}{\nu} = \frac{\Delta Cd}{1-Cd} \tag{8}$$

$$\frac{\tau}{\delta} u^2 + u + ke^{-u} = 0 \tag{9}$$

The range of communication delay which gives the real root of the characteristic equation of formula (9) is narrower than the case of no C_d (Fig. 4(a)). It means that the C_d is an instability factor to the target approaching action (Fig.4 (b)).

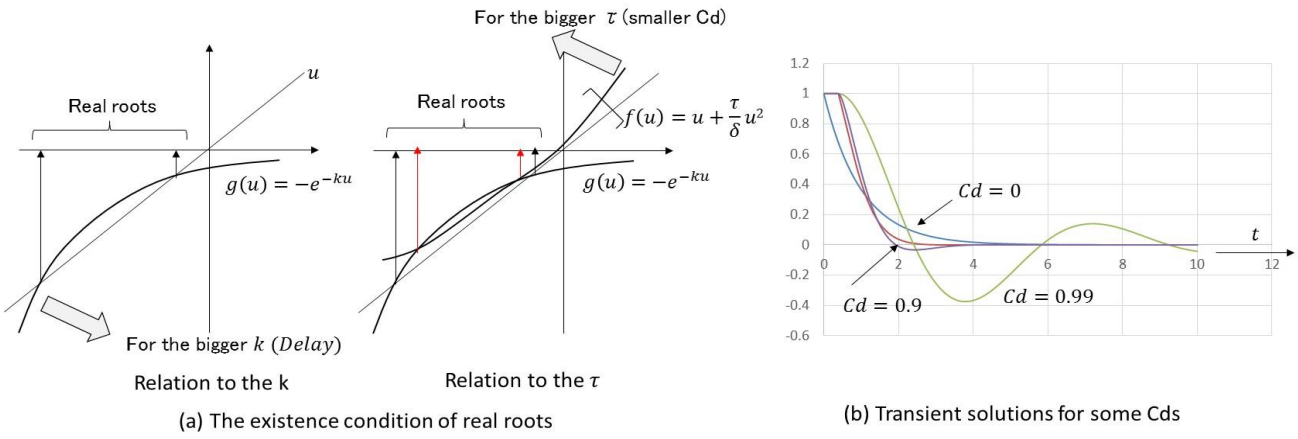


Fig. 4: The behaviour of solution of (5)

4. The Effect to the Reaction Force Feedback

For the feedback of reaction force from slave to master, it is necessary to transform the reaction force to some displacement at the master. The representative means is to use the spring dumper mechanism in the master's interface. In this case, the feedback loop of the reaction force is described as the control brock diagram as shown in Fig. 5 [12]. The Equation of motion of the system shown in Fig. 5 with communication delay δ is described as formula (10).

$$\frac{m}{c}\frac{d^2y(t)}{dt^2} + \frac{dy(t)}{dt} + \frac{\kappa}{c}y(t-\delta) = 0 \tag{10}$$

where m is the mass, κ is the spring constant and c is the dumper constant.

Introducing the resistance to this system, the difference differential equation and the characteristic equation are obtained as formulae (11) and (12).

$$m\frac{d^3y(t)}{dt^3} + \left(c + \frac{m}{\tau}\right)\frac{d^2y(t)}{dt^2} + \frac{c}{\tau}\frac{dy(t)}{dt} + \frac{\kappa}{\tau}y(t-\delta) = 0 \tag{11}$$

$$m\frac{d^3y(t)}{dt^3} + \left(c + \frac{m}{\tau}\right)\frac{d^2y(t)}{dt^2} + \frac{c}{\tau}\frac{dy(t)}{dt} + \frac{\kappa}{\tau}y(t-\delta) = 0 \tag{12}$$

Introducing big C_d means big τ. Examing the characteristic equation (12), it is concluded that the bigger C_d contributes to the system stability (Fig.6).

5. Towards the Optimization of C_d

From the above theoretical analysis, the resistance C_d has the effect to decreasing mechanical howling but it is the degrading factor to the target approaching action. Therefore, the one of the methods to optimize the value of C_d is to determine as the minimum value not to cause the mechanical howling. To estimate the system parameters from the experiment and to obtain the minimum C_d from the examination of the characteristic equation is a possible way. To say concretely, the stability condition is that the characteristic equation has only real roots.

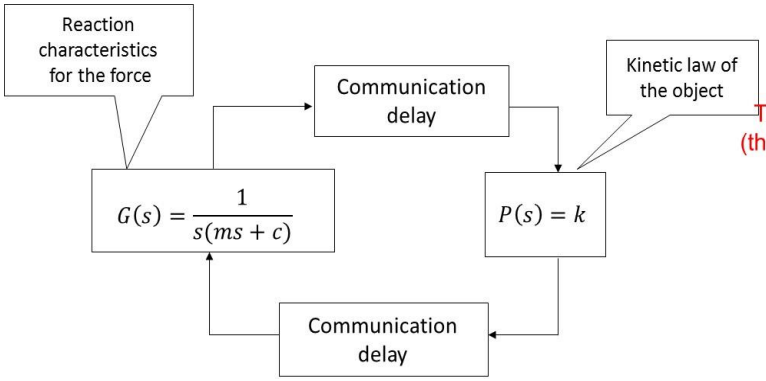


Fig. 5: The brock diagram of the force feedback

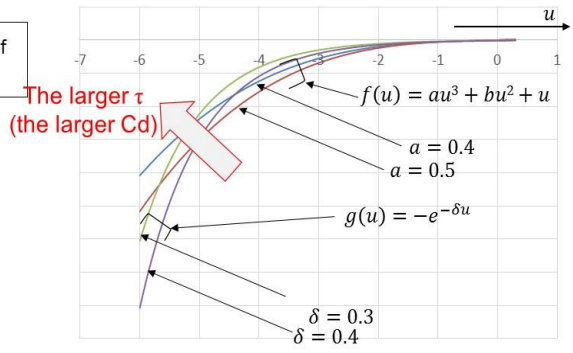


Fig. 6: Examination of the real roots of equation (12)

Although these analytical methods are imagined, firstly it is necessary to understand human characteristics. There are two human characteristics according to the two feedback loops. Of course the target approaching action depends on the human characteristics, but also the force feedback depends on the human characteristics. Because the element of kinetic parameter such as the dumper parameter contains the operator's physical characteristics.

6. Human Characteristics of Target Approaching Action

The target approaching human characteristics is returned to the function described in formula (2). We have tried to grasp human characteristics by experiment using the distributed game system with haptic communication [13], [8]. The system is a game in which players move an object on a virtual space. The objects are moved along the maze and the maze can be constructed by using the movable walls. The players move the objects to the settled goal (Fig. 7). The communication delay can be changed.

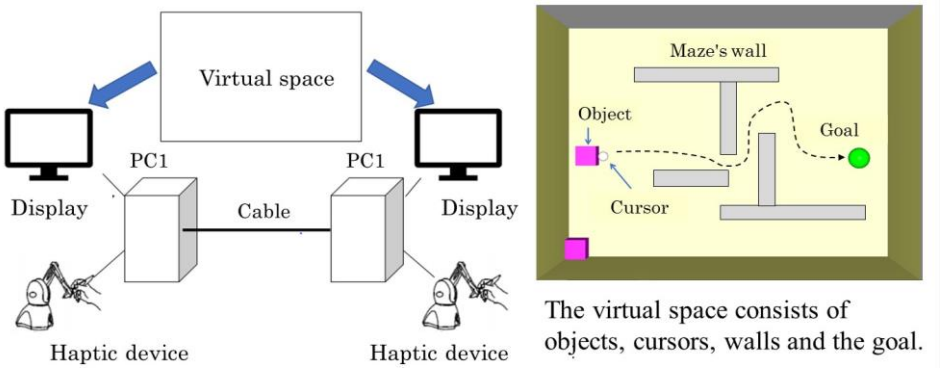


Fig. 7: System configuration of developed game

By using the system, the following analysis becomes to be possible. (a) Play not only in the simple route, but also in various complex routes. (b) Not only the standalone work, but also the competitive or cooperative work can be analysed. (c) The effect of communication delay can be investigated.

The obtained data from this system is the chronological ordered data of time and position (t, y(t)). The parameter A and α is estimated by this data. For estimating these parameters from the data, it is thought to be expressed as the scatter plot of (y, V). However, the destination point exists only in the operator's head and cannot be measured from outside data. Therefore, we have proposed the following method. The derivation of formula (3) under the assumption that A and α are constant yields formula (13).

$$log\left(\frac{d^2y(t)}{dt^2}\right) - log\left(\frac{dy(t-\delta)}{dt}\right) = \left(\frac{1}{\alpha}\right) logA + log\alpha + \left(1-\frac{1}{\alpha}\right) log\frac{dy(t)}{dt} \tag{13}$$

Formula (13) means that $ln(d^2y/dt^2)$ is a linear function of $ln(dy/dt)$. Therefore, from the linear regression of the scattered plot of $ln(d^2y/dt^2)$ and $ln(dy/dt))$, A and α can be obtained. The followings are the estimation example. Fig. 8(a) is the moving route of the object, Fig. 8(b) is the relationship of position and velocity, and Fig. 8(c) and Fig. 8(d) are the estimation results of A and α for different delay conditions.

The velocity functions in Fig. 8(c) and (d) are different from each other. The theoretical analysis using the estimated parameter under no delay says that the actual action has smaller vibration than predicted one by theoretical analysis. This is thought that the operator might absorb the effect of communication delay unconsciously and it is appeared to the difference of velocity function. Therefore, the difference of velocity function might be a large factor of QoE of the haptic systems.

7. Experiment of the Mechanical Howling

The human characteristics related to the mechanical howling has been analysed by experiment. Moreover, the estimation method of the velocity function has also been improved.

7.1. Improvement of the Estimation Method

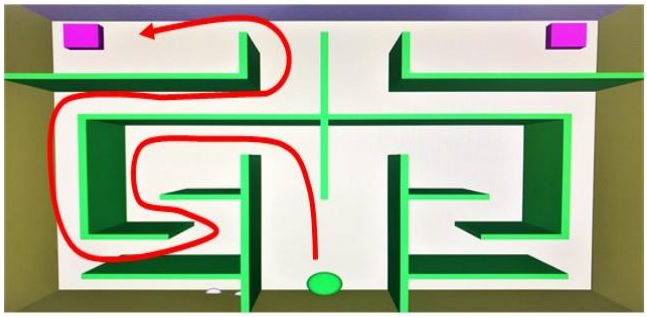
The observed positon $y(t)$ is different form the master's instructed position if the C_d is introduced. Let $m(t)$ be the master's instructed position. In this case, the velocity function is $dm(t)/dt$ in formula (14).

$$\frac{dm(t)}{dt} = -A(y(t-\delta) - y_D)^{\alpha} \tag{14}$$

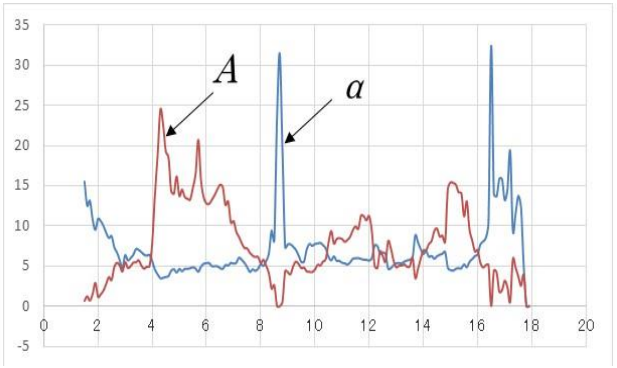
Firstly, get *m(t)* from *y(t)* by using formula (1) and analyse the chronological ordered data of time and position (t, y(t)). The deviation of formula (14) yields the following formula.

$$log\left(\frac{d^2\boldsymbol{m}(t)}{dt^2}\right) - log\left(\frac{dy(t)}{dt}\right) = \left(\frac{1}{\alpha}\right) logA + log\alpha + \left(1 - \frac{1}{\alpha}\right) log\frac{dm(t)}{dt} \tag{15}$$

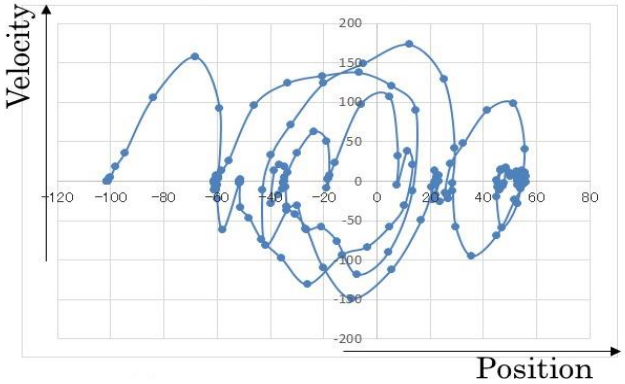
After here, the estimation method is same as mentioned in section 6. However, estimation is not done when the contribution rate in the current window is less than the certain value. In such a case, the velocity changes with rapidly like vibration.



(a) Moving route



(c) Velocity in on delay



(b) Position and velocity

(d) Velocity in communication delay

Fig. 8: Estimation example of velocity function

7.2. Estimation Examples

The experiment is carried out by the route shown in Fig. 9. Object is moved from the start point and pass the narrow passage and hit against the wall. Fig. 10 is an example of the 2-dimension trajectory. The part circled in the figure means the cause of mechanical howling. With or without of mechanical howling under some condition of C_d and communication delay are shown in Fig.11. When C_d is large the mechanical howling is suppressed even the case of large communication delay. This is consistent to our theory.

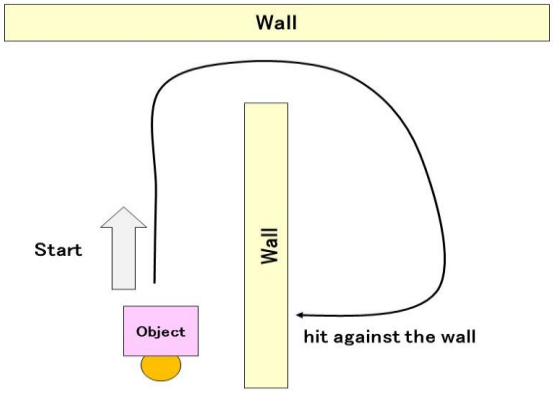


Fig. 9: The experiment route

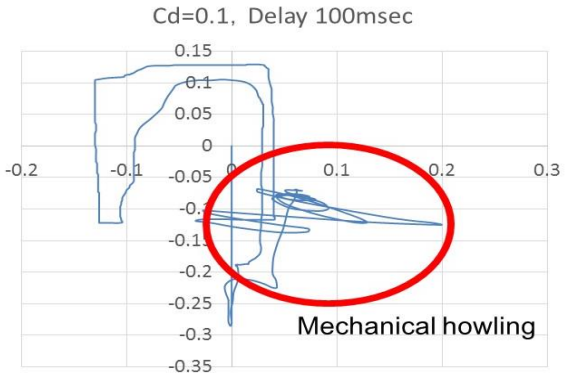


Fig. 10: A example of trajectory

Cd	Delay (msec)				
	0	50	100	200	300
0.1	Severe	Slightly	Severe	Slightly	Slightly
0.3	No	—	—	—	Slightly
0.5	Slightly	—	—	—	Severe
0.7	Slightly	—	—	—	No
0.9	No	No	No	No	Slightly

Fig. 11: With or without of mechanical howling

Fig. 12 shows the temporal change of position and velocity function of a specific player. The temporal change of position is greatly different according to the condition of *Cd* and communication delay. However, the velocity function hardly change. This is different from the result mentioned in section 6. It is thought that the work condition might affect to the player's operation. The maze of this experiment is not complex as section 6. The results in former section shows that the solution of difference differential equation changes according to delay and C_d if the velocity function is same. Therefore, the change of positon is great and the change of velocity function is small is consistent to our theory.

Fig. 13 shows the individual differences of velocity function under similar condition. The individual differences are large. Player 3 has smaller α than other player. The small α means the rapid acceleration deceleration. This characteristics of player 3 reflects the temporal change of the positon.

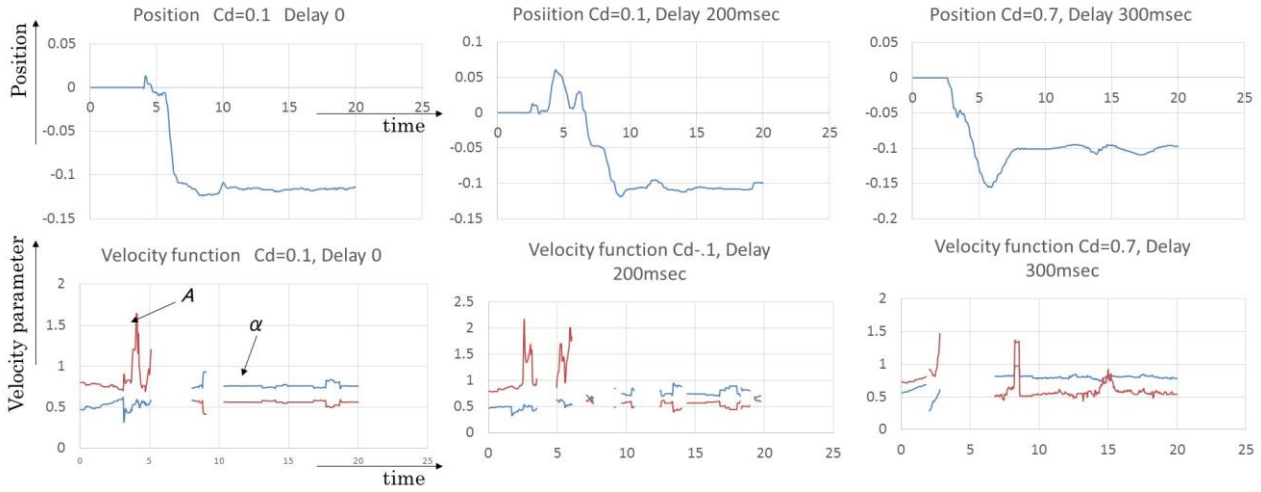


Fig. 12: Change in Velocity Function

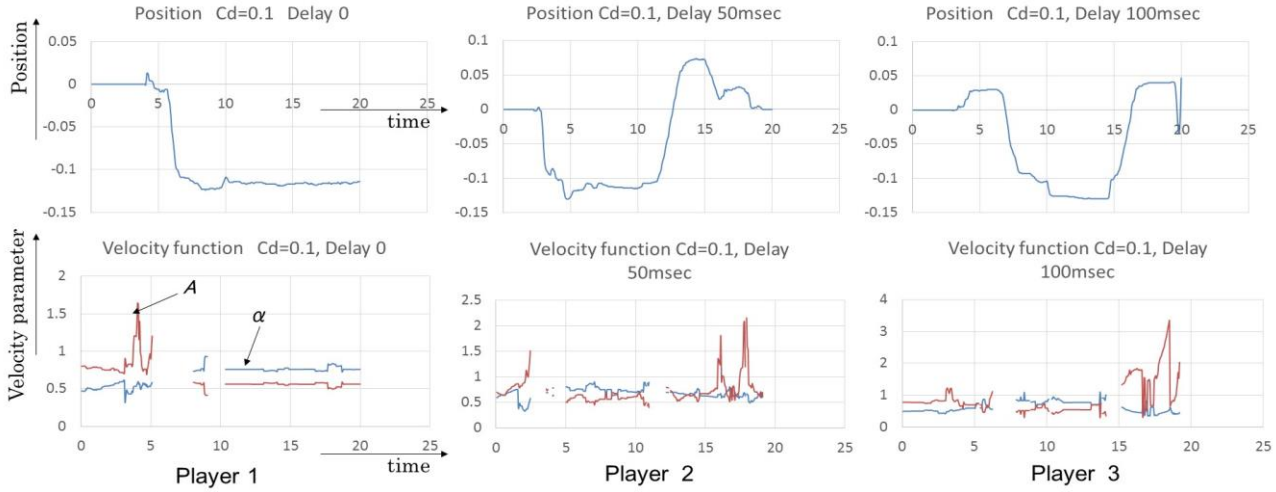


Fig. 13: Individual difference of velocity function

8. Conclusion

This paper investigated the feasibility of the determination of optimum resistance to satisfy both stability and operability. Our previous studies have been summarized and the new results have been presented. From the experiment, including our previous studies, operators adjust their motion under the communication delay to decrease the vibrations. But these phenomena are depend on many conditions such as kind of work, complexity of work and so on. Furthermore, the individual different of velocity function is large. The final purpose of this research is to realize the good operability and the stability. Generally speaking, Cd is a kind of signal conversion and to find the optimal conversion is the future problem. To solve this problem, the feedback for approaching the target position and the feedback of reaction force should be integrated and more investigation to clarify the operating characteristics is required.

9. Acknowledgements

This work was partly supported by JSPS KAKENHI Grant Number 18K11261.

10.References

[1] K. Ohnishi, “Real world haptics : Its principle and future prospects,” The journal of the Institute of Electrical Engineers of Japan (IEEJ), vol. 133, no. 5, pp.268-269, Mar. 2013.

[2] T. Kawai, “Haptics and surgery,” IEEJ Journal, vol. 133, no. 5, pp. 282-285, Mar. 2013.

[3] ITU-T Rec. E. 800, “Terms and definitions related to quality of service and network performance including dependability,” 1994.

[4] ITU-T Rec. G. 100/P. 10 Amendment 1. New appendix I - Definition of quality of experience (QoE), 2007

[5] Y. Komatsu, H. Ohnishi, and Y. Ishibashi, “Adaptive control of viscosity in remote control system with force feedback,” in Proc. IEEE International Conference on Consumer Electronics - Taiwan (ICCE-TW), pp. 237-238, June 2017.

[6] H. Watanabe, Y. Ishibashi, and P. Huang, "A formulation of remote robot system by using difference differential equation," in Proc. IEEE The 3rd International Conference on Computer and Communication Systems (ICCCS), pp. 511-515, Apr. 2018.

[7] H. Watanabe, P. Huang, and Y. Ishibashi, "An investigation of the stabilization of bilateral robot systems under communication delay," in Proc. IEEE International Conference on Intelligence and Safety for Robotics (ISR), pp. 140-145, Aug. 2018.

[8] H. Watanabe, K. Kuroyanagi, Z. Sato, H. Hirado, P. HUANG and Y. Ishibashi, " A Proposal of the Method for Analyzing the Stability of Virtual Distributed Systems using Haptic Communication," in Proc. IEEE International Conference on Consumer Electronics - Taiwan (ICCE-TW), pp. 55-56, May 2018.

[9] S. Sugiyama, “Difference differential equation,” Kyoritsu-shuppan, 1971.

[10] R. Miyazaki, “Introduction to the theory of delay differential equations,” Report of the Research Institute for Mathematical Sciences, Kyoto University, published in 2010.

[11] H. Watanabe, Y. Ishibashi, and P. Huang," A Stability Analysis of Haptic Systems by Using Difference Differential Equation, -From the view point of characteristic equation- " ,in Proc. The 2nd World Symposium on Communication Engineering (WSCE 2019), pp. 127-132, Dec. 2019.

[12] T. Miyoshi, “Stabilized Control in Haptic Communications”, J. IEICE, Vol.102 No.1 pp.57-63, 2019.

[13] Pingguo Huang, Yutaka Ishibashi, “QoE assessment of will transmission using vision and haptics in networked virtual environment”, International Journal of Communications, Network and System Sciences (IJCNS), vol. 7, no. 8, pp. 265-278, Aug. 2014.)

ISBN 978-981-14-4787-7
Proceedings of 2020 the 10th International Workshop on Computer Science and Engineering (WCSE 2020)
Yangon (Rangoon), Myanmar (Burma), February 26- February 28, 2020, pp. 122-129

Artificial Intelligence based 6G Intelligent IOT: Unfolding an Analytical Concept for Future Hybrid Communication Systems

Muhammad Haroon Siddiqui [1+], Kiran Khurshid [1], Imran Rashid[1] and Adnan Ahmed Khan [2]

[1] Electrical Engineering Department, National University of Sciences and Technology, Islamabad, Pakistan.

[2] Computer Software Engineering Department, National University of Sciences and Technology, Islamabad, Pakistan.

Abstract. The Internet of Things (IOT) is a new paradigm for wireless communication networks with access to artificial intelligence (AI). The conventional cellular network is experienced with two major issues such as spectrum scarcity and insufficient intelligent autonomous competency. These problems are required to be addressed otherwise it will obstruct the massive deployment of next generation applications. The concept of the next generation communication network, creates the opportunity to design a network with number of sensors in the paradigm of IOT to process large data. In this paper, we present a novel 6G intelligent IOT paradigm to optimize communication channels and process big data intelligently. Firstly, we present the proposed architecture of 6G intelligent IOT paradigm, then we discuss the enabling technologies in relation to our proposed concept. In 6G scenario, AI is not very effective as compared to the physical layer (PHY), however, it provides flexibility and agility to the air interface with enhanced efficiency. We discuss AI based 6G air interface architecture, based on our proposed paradigm. Finally, we evaluate the results through simulation by comparing our propose paradigm with 5G-IOT and 5G intelligent IOT. Simulation results confirm that our proposed paradigm out performs the others.

Keywords: Artificial intelligence, cognition cycle, deep learning, super massive-MIMO, 6G.

1. Introduction

5G wireless network is considered, a key enabler, for intelligent information society of 2020. Extensive efforts are being made by 3rd generation partnership project (3GPP) to encourage further development of 5G technologies. In the meantime, IEEE 8011ax standard for wireless local area network is being introduced [1]. It is anticipated that even after introduction of millimeter wave (mm-wave) and massive multi input multi output (MIMO) (with large scale arrays of antennas), 5G can maximum achieve 20 Gb/s transmission for end user. The latest applications such as internet of things (IoT), artificial intelligence (AI), cognition cycle (CC), IP multimedia subsystem (IMS), vehicular network and machine learning, will depict the next generation of cellular network, viz: 6G. Investigations and research on 6G wireless networks has also been made a part of agenda to meet the needs for the intelligent information society 2030 [2], [3]. China has also introduced a plan "broadband communication and new networks" for 2030 and beyond. The university of Oulu (Finland), has launched 6G research program viz: 6G-enabled wireless smart society and ecosystem (6Genesis) [4], [5], which will explore the avenues for the development of 6G standards. Various projects related to beyond-5G (B5G), have been sponsored by european commission's horizon 2020 program. In United States, federal communication commission (FCC) has already started research on 6G and launched THz band. Intelligent information society 2030 is envisaged highly digitized, intelligent, autonomous, globally data driven and unlimited wireless connectivity [6]. Fig.1 shows the resource utilization of 6G based on time-frequency-

[+] Muhammad Haroon Siddiqui. Tel.: + 923335125126.
E-mail address: haroon@mcs.edu.pk

space. The concept of 6G will be the key enabler in order to achieve this blueprint. It is anticipated that 6G will be autonomous with intelligence and consciousness like human. Moreover, 6G technology will:

- Support ultra-high definition videos with extremely high throughput
- Allow highly low latency communications such as 10 μs, especially for industrial internet [7]
- Support Nano technologies which include internet of nano-things, implantable nano-sensors and nano-devices with extremely high energy efficiency
- Support space and deep sea communications
- Enhance and support existing 5G key applications.

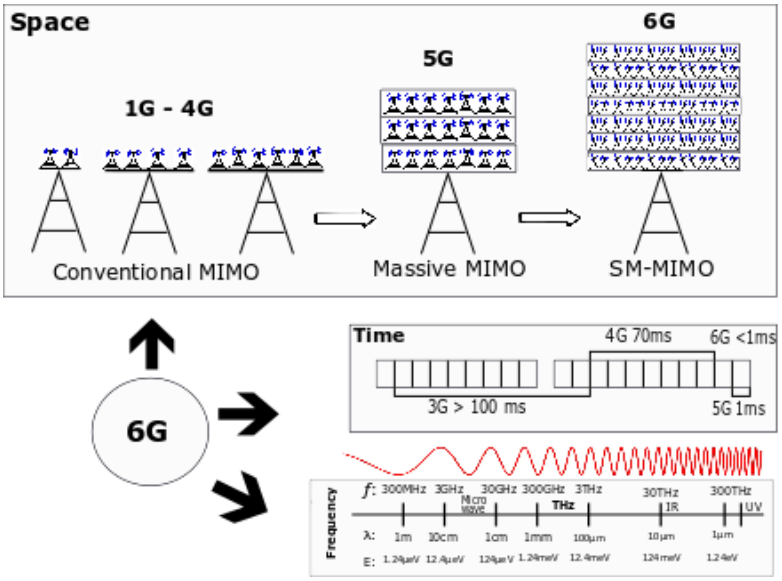


Fig. 1: An elucidation of resource utilization of 6G based on time-frequency space.

The ever growing attraction especially in the domain of 6G, IoT, AI and big data, draws the attention of researchers to fully comprehend their conceptual structure, future potential and challenges. Likewise, the future concept of communication systems with enormous sensors, demands convergence of key technologies to produce novel paradigm. 6G intelligent IoT paradigm, is the convergence of intelligence, internet and things. The concept of 6G is a backbone for realization of future IoT technology as IoT applications require to handle a large number of heterogeneous devices. A network of such ever-increasing number of devices can benefit from 6G technology. 6G technology can provide the connectivity, desired network standards and architecture for real time processing of the massive data with low latency rate. In this paper, we make an endeavour to propose an architecture of 6G intelligent IoT and evaluate its major components. Then we introduce some of the key and enabling technologies for 6G intelligent IoT with their application on proposed paradigm. Finally, we evaluate the performance of proposed paradigm on the basis of some performance metrics.

2. The Architecture of 6G Intelligent IoT

Fig.2 shows the architecture of 6G intelligent IoT, which is based on existing cellular network. It is a general cellular network with the latest emerging technologies, such as, super massive (SM)-MIMO, ultra-dense static and mobile cells networks [8]. It also presents the BSs and a cloud in 6G intelligent IoT paradigm. The concept of device-to-device (D2D) communication [9] and small cell access points are also integrated in this architecture. 6G intelligent IoT paradigm comprises of three major components.

2.1. Intelligent Ultra High Speed Processing Core (IUSPC)

6G intelligent IoT paradigm is envisaged for autonomous, intelligent and without human intervention applications. To make this component more intelligent and avoid repetition to reduce latency time, cognition cycle (CC) is used for learning and previous memory. AI techniques, such as swarm intelligence, genetic algorithm, reinforcement learning, and multi-layer perceptron are being used to implement CC [10]. The

other two components in the cloud include intelligent computing module (ICM) and execution module (EM). ICM comprises of several intelligent computing system and known as the nerve center. AI technique (deep learning) is used for the online learning realization. In all ICM, multiple servers are grouped together for intelligently processing of massive data. This arrangement allows ICM to operate independently and efficiently. ICM is also responsible to process the data intelligently from the Intelligent Base station (IBS) and passing the instructions to the IBS accordingly. CC plays a vital role in the scheme in order to reduce the latency rate as learning from the past experiences, save the processing time required for computation.

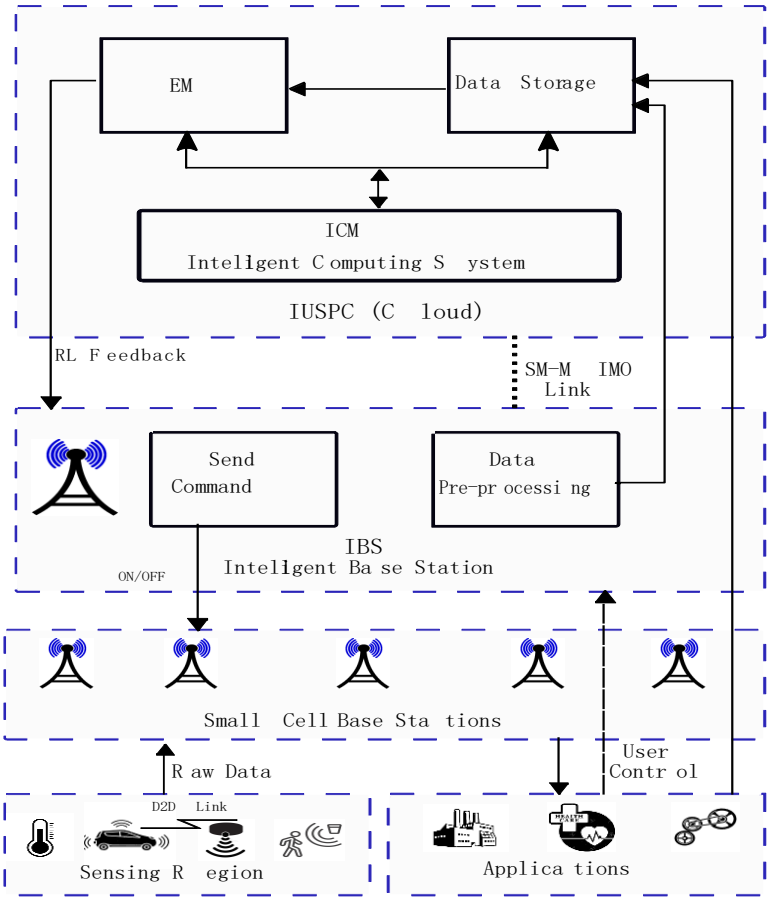


Fig. 2: 6G intelligent IoT architecture.

2.2. IBSs

An IBS is consisted with all requisite requirements to connect IUSPC with the sensing regions. Raw data is processed at IBS, then processed data is transmitted to IUSPC. On the basis of feedback received from IUSPC, more intelligent schemes are implemented. It is anticipated that in future communications especially in heterogeneous operating environment, massive unprocessed data, transmitted to the BS through ultra-dense small cell networks, is required to be processed at high cost. Hence, unprocessed data must be pre-processed redundantly. The IBS receives the instructions from the IUSPC for intelligently execution, however, if the channel is disconnected the sensing region will stop transmission. CC, is embedded in IBS and helps in selecting (through learning from past experience) the correct unoccupied channel at right time to prevent unnecessary transmission. This will reduce the traffic load, enhance the data processing speed and improve the energy efficiency.

2.3. Sensing Region

This region has sensors in large numbers, which are used to compute or determine the physical quantity and convert it into digital signal for processing. The 6G terahertz (THz) will support more effectively the emerging applications such as D2D communication [11] and internet of vehicle (IoV) [12]. These applications emerge in the physical world or sensing region. The main function of the sensing region is to observe the state of particular domain. On the basis of the nature of assignment, these regions are divided into mobile sensors region (deals with mobile devices) and fixed or static sensors region. These regions assist

IoT devices to interact with the environment and collect the sensing information. This information is subsequently transmitted to THz small cell networks and then to IBS for further processing in 6G environment.

3. Promising Technologies to Implement 6G Architecture

3.1. Message Queuing Telemetry Transport (MQTT) Protocol

In ubiquitous communication environment, IoT devices communicate through MQTT protocol [13]. The MQTT uses TCP/IP for network connections and can send simultaneous messages to multiple receivers. In an intelligent IoT based 6G architecture, MQTT protocol is used for interconnection of the various components [14]. MQTT protocol has some inherent qualities due to which the chances of loss and duplication of the data information is reduced, hence, results in efficient use of the radio resources.

3.2. THz

THz communications [15] band (in the range of 0.1−10 THz) uses richer spectrum resources than that of mm-wave band in 5G. It can take the benefits of both light and electromagnetic waves. In 6G architecture, THz communication is expected to provide multi-Tb/s data transmission to various emerging applications at sensing region [16]. Moreover, the benefits, anticipated, to achieve through THz communication are:

- It can support massive bandwidth demand as envisage in 6G technology and provide multi-Tb/s transmission.
- Due to shorter wave length, THz can integrate multiple antennas, hence can provide thousands of beams. It is anticipated that 10000 antennas can be accommodated in THz based IBSs which results in achieving extreme narrow beams and ultra-high data transmission.
- Inter-cell interference free and secure communication due to highly directional transmission.

3.3. SM-MIMO, Large Intelligent Surface and Holographic Beamforming

MIMO, with more than ten thousand of antennas, enhances the capacity manifold and provides reliable transmission by diversity [17]. It also reduces the propagation losses by narrow beamforming. In a 6G intelligent IoT paradigm, SM-MIMO is expected to offer following benefits:

- It improves energy efficiency and enhances the latency rate.
- Increases network throughput.
- Massive access communication can be achieved by combining SM-MIMO and non-orthogonal multiple access (NOMA) schemes.
- Ultra-narrow beams reduce propagation losses as well as lessen the co-channel inter-cell interference.

3.4. Big Data Mining

In 6G Intelligent IoT paradigm, real time data processing is required, hence, big data generated in this scenario is altogether different from conventional data. Data collection signals from small cells is broadcasted to the network management. Data is collected from both mobile and fixed sensor regions according to the instructions. Then raw data is sent to the IBSs where data is pre-processed. The methods used for pre-processing, include integration, cleansing and redundancy. Data integration gathers data from all the sources and combines it. In cleansing, irrelevant, incomplete and unwanted data is deleted. Finally, in redundancy, redundant data is removed and compression is used for efficient utilization of spectral resources.

3.5. Deep Learning

Deep learning is a branch of machine learning, which is based on learning of data representations from different architectures. Deep learning has already been applied to many applications, which has proved to be an efficient technique. It is regarded as a key step toward actual AI. Massive data is produced with the development of intelligent IoT in a 6G paradigm. Hence, big data mining and deep learning can render very effective and efficient services in this scenario. Fig.2 also shows that ICM and EM is placed at the center, so that data can be processed automatically with intelligent algorithms.

Various independent intelligent computing systems are designed to process unique data intelligently, such as a vehicle number plate detection system, a human face identification system, and any specific oddity detection system. They implement deep learning algorithms and act within specific scenario as an ICM. Every ICM is a unique system which processes distinct data, where the processed data information is readily available to be shared among systems. ICMs collect real time data from various sources, process it, and forward it to EM for further disposal before storing it locally or being sent it to the users or in the cloud.

3.6. Reinforcement Learning

Reinforcement learning presents the scenario of the states in a particular environment by taking various actions and receiving rewards. One of the key feature of reinforcement learning is to find the optimal solution by investigating an unknown environment. The major three components in this perspective involved in the procedure are: agents, rewards and policies. In a 6G intelligent IoT paradigm, an agent is described as the collection of IBSs and IUSPC. Data from all the sensing regions are fed as an input, whereas, channel's states activity from the EM, is controlled by the command which is the output of an agent. In a big data environment, reward is generated for each command. To ensure the quality of experience (QoE) of users, CC keeps the track of channel occupancy.

A policy is formulated, with corresponding control command for implementation, at IUSPC. The essence of each policy, is to provide an efficient and accurate judgment. After successful evaluation process, policy is introduced to the communication system. In an EM, multiple probability coefficients are defined according to the various applications scenarios. The probability coefficients specify the significance of each ICM. After initialization of probability coefficients and processing of the data at ICM, processed data, is multiplied by their corresponding coefficients. Then the result is compared with predesignated threshold in EM. If the resulting quantity exceeds the threshold, then a control command is sent to IBS for closing the transmission and restricting other unwanted data to upload.

Reinforce learning technique, assists EM in automatic learning of the probability coefficients and thresholds for various application as well as intelligently online execution. Likewise, ICM can also learn automatically in various applications environment and can process the data online, intelligently and accurately. It has been experienced that through online techniques, automatic and intelligent performance can be achieved, which is definitely far better than the manual adjustment of threshold.

3.7. CC

CC is the concept of learning from the previous actions for optimal outcomes with the passage of time. There are basic five steps are involved in implementing CC, Fig.3 shows various stages of CC [18], [19]. In first stage data is collected from sensor regions and is observed thoroughly. After observation stage, received data is analyzed to identify the pattern. In third stage, course of action is adopted from various options. In fourth step, actions are executed. Lastly, it learns the knowledge from the outcome of executed actions for future compliance.

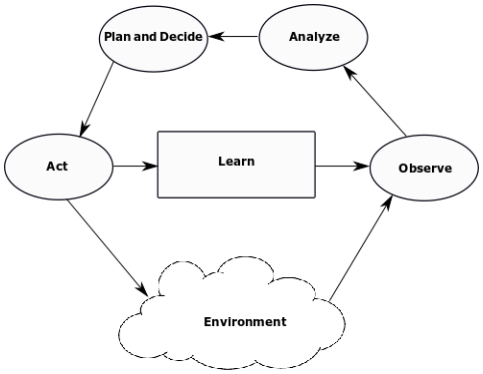


Fig. 3: An illustration of cognition cycle.

4. AI Based 6G Interface

Table-I explains the air interface implementation for 6G intelligent IoT paradigm. AI will notably revolutionize the architect of 6G air interface.

Table 1: Ai Impacts on 6g Network Functions

	Functions	AI Algorithms	Descriptions
Physical Layer	- Reliable Data Transmission Channel coding, modulation, MIMO precoding and OFDM - Channel Estimation	- DNNs - Autoencoder - CNNs - CCNNs - K-Means	- Revamp end-to-end PHY architecher reduces complexity level of MIMO OFDM receiver - During high mobility enhances PHY performance
Data-Link Layer	- Frame flowing operations flow control synchronization, data packets queing, scheduling power control, error correction and flow control etc	- DNNs - Supervised learning - Deep learning - Reinforcement Learning - Transfer Learning - Q- Learning	- Optimal user scheduling Improve network performance Channel Estimation Traffice Prediction Enhances radio resource efficiency - Optimal retransmission redundancy - Reduction in retransmission overhead
Network Layer	- Responsible for management of: Riuting,Mobility,RRC connection and load - BS association - BS clustring	- DNNs - K-Means - Unsupervised Learning - Supervised learning - Reinforcement Learning - Q- Learning	- Optimizes serving cells - Optimizes multiple connectivity - Mobility prediction - Optimizes handover - Optimizes data transmission paths - Optimizes BS clustring - Controlling size of cluster in dynamic network scenerios

4.1. Physical-Layer (PHY)

It is the most significant layer in achieving the reliable and rapid data transmission over wireless channels. At transmitter side, various modules perform operations such as coding, modulation, orthogonal frequency division multiplexing (OFDM) and MIMO precoding. Likewise, reverse operations are performed at receiver side to recover the original data. In 6G intelligent IoT paradigm, an AI based end-to-end architecture is designed, in which auto-encoders from deep learning can be used to perform processing of the data intelligently. However, to avoid complexity and make this scheme more practical, AI application is designed independently which enhances the functions of PHY. In deep learning method, convolution neural networks (CNNs) can perform signal classification and channel decoding, whereas OFDM is built with complex CNNs (CCNNs). Signal detection and channel estimation can be performed through deep neural networks (DNNs).

4.2. Data-Link Layer

AI can enhance the sub-layers of data-link layers including medium-access control, packet data-convergence protocol, radio-link control and service data adaptation protocol. AI enabled resource allocation, is capable of selecting the most appropriate scheduling for users with (or without) traffic prediction [20]. Security enhancements can be achieved by adopting suitable AI security algorithms [21]. In intelligent IoT 6G scenario AI security algorithms are very effective due to the short packets. Another key feature of AI, is to optimize the operation of automatic repeat request (ARQ) and hybrid ARQ. It can lessen the re-transmission overheads and improve the reliability of data transmission.

4.3. Network Layer

Network layer generally renders IBS and user specific functions which include traffic load balancing, radio resource control (RRC) and mobility management. AI techniques enable user to choose optimal serving cells, multiple connectivity and various other features, which can guarantee the continuity of service. Likewise, with AI, IBSs can balance the traffic load and improve the network robustness.

5. Performance Evaluation and Analysis

The basic aim of an 6G intelligent IoT paradigm is to achieve intelligently big data processing and enhance the efficiency of communication channels to optimum. There are quite number of performance evaluation metrics, to evaluate the actual performance of a 6G communication networks, such as complexity implementation, spectral efficiency, energy efficiency, latency, data rate, QoS, processing speed and many more. It is very hard to incorporate all the performance metrics to evaluate the efficacy of 6G intelligent IoT due to high complexity level. However, we chose effective utilization of channels (EUOC) as a performance indicator to assess our simulation results. The EUOC is computed by k/m, where k is the group of various service data and m is the group of valid service data. Simulations of 5G IoT, 5G intelligent IoT and 6G

intelligent IoT are shown in figures with blue, red and black line respectively. In Fig.4, EUOC performances of 5G IoT and 5G intelligent IoT, are compared. Result shows that 5G intelligent IoT performs better than 5G IoT. Likewise, Fig.5 exhibits the performances of 5G IoT and 6G intelligent IoT paradigm. It is very evident from the simulation results that proposed 6G intelligent IoT outperforms the 5G IoT in terms of EUOC performance. In Fig.6, the EUOC performances of 5G intelligent IoT and 6G intelligent IoT are compared. Simulation results confirm that the proposed architecture shows significantly better results as compared to 5G intelligent IoT. In Fig.7, EUOC performances are compared for three scenarios, including 5G IoT, 5G intelligent IOT and 6G intelligent IoT. Experimental results show that performance of 6G intelligent IoT is the most effective.

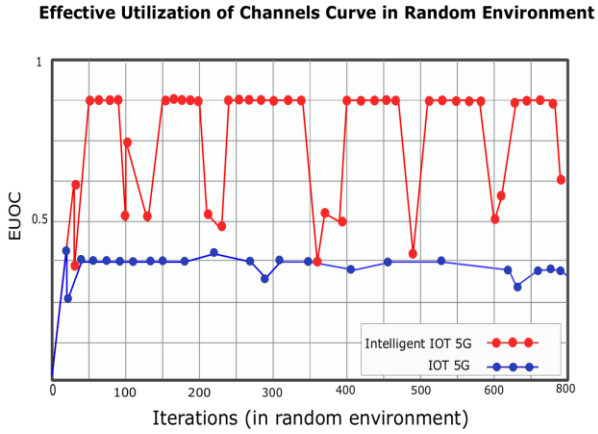


Fig. 4: Performance evaluation curves of effective utilization of channels (EUOC).

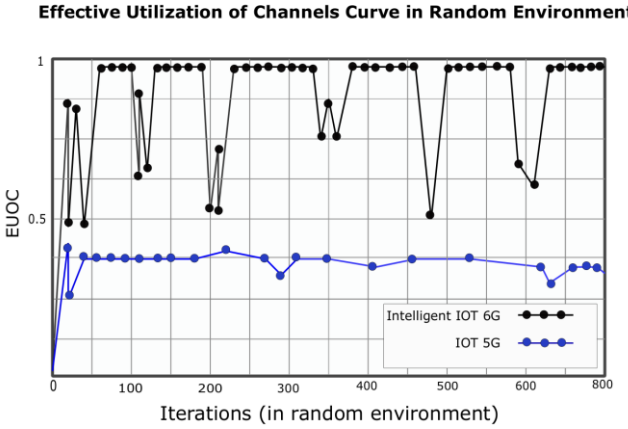


Fig. 5: Performance evaluation curves of effective utilization of channels (EUOC).

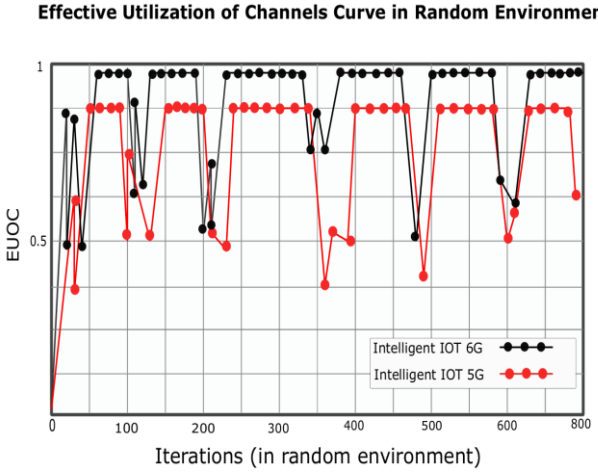


Fig. 6: Performance evaluation curves of effective utilization of channels (EUOC).

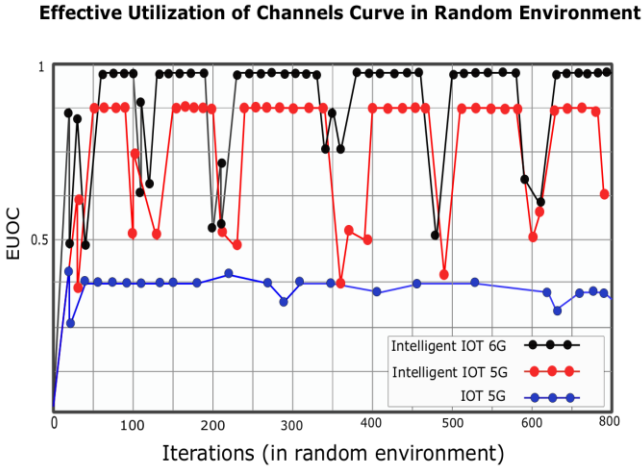


Fig. 7: Performance evaluation curves of effective utilization of channels (EUOC).

6. Conclusion

We have presented a novel concept of the 6G intelligent IoT, which converges 6G, IoT and emerging technologies to an innovative paradigm. Our proposed architecture, can cater to the ever increasing demand of future communication network, especially in the domain of optimization of communication channels and processing of massive data. First, we have presented the three main elements of 6G intelligent IoT paradigm viz: IUSPC, IBS and sensing regions. Then we have introduced some enabling technologies with application to 6G intelligent IoT, such as MQTT protocol, THz band, CC, big data mining, deep learning and reinforcement learning. AI based air interface has also been presented with an application to the 6G intelligent IoT framework. Finally, we have evaluated the experimental results of proposed architecture on the basis of performance indicators. We hope that this work gives an innovative concept on IoT, in which emerging technologies are applied to the collaborative paradigm of 6G and IoT.

7. References

[1] E. Khorov, A. Kiryanov, A. Lyakhov, and G. Bianchi, ``A tutorial on IEEE 802.11ax high efficiency WLANs", IEEE Commun. Surv. Tut. ,vol. 21, no. 1, pp. 197–216, 1st Quarter 2019.

[2] K. David and H. Berndt, ``6G vision and requirements: Is there any need for beyond 5G?", IEEE Veh. Technol. Mag., vol. 13, no. 3, pp. 72–80, Sept. 2018.

[3] R. Li, ``Network 2030: Market drivers and prospects" In Proc. 1st International Telecommunication Union Workshop on Network 2030, Oct. 2018. [Online]. Available: https://www.itu.int/en/ITU-T/Workshops-and-Seminars/201810/Documents/Richard_Li_Presentation.pdf.

[4] A. Pouttu, ``6Genesis–Taking the first steps towards 6G", In Proc. IEEE Conf. Standards Communications and Networking, 2018. [Online]. Available: cscn2018.ieee-cscn.org/files/2018/11/AriPouttu.pdf.

[5] Z. Baiqing, Z. Xiaohong, W. Jianli, L. Xiaotong, and Z. Senlin, ``Photonics defined radio–A new paradigm for future mobile communication of B5G/6G", In Proc.6th Int. Conf. Photonics, Optics and Laser Technology, pp. 155–159, 2018.

[6] M. Latva-aho, ``Radio access networking challenges towards 2030", 1st International Telecommunication Union Workshop on Network 2030, Oct. 2018. [Online]. Available: https://www.itu.int/ en/ITU-T/Workshops-and-Seminars/201810/Documents/Matt_Latvaaho_Presentation.pdf.

[7] ``Unified Architecture for Machine Learning in 5G and Future Networks", International Telecommunication Union–Telecommunication Standardization Sector, Technical Specification ITU-T FG-ML5G-ARC5G, Jan. 2019.

[8] S. H. Shah and I. Yaqoob, ``A Survey: Internet of Things (IOT) Technologies, Applications and Challenges", Smart Energy Grid Engineering, pp. 381–85, Aug. 2016.

[9] L. D. Xu, W. He, and S. Li, ``Internet of Things in Industries: A Survey", IEEE Trans. Industrial Informatics, vol. 10, pp. 2233–43, Jan. 2014.

[10] J.Qadir, ``Artificial intelligence based cognitive routing for cognitive radio networks", Artif. Intell. Rev., vol. 45, no. 1, pp. 25–96, Jan. 2016.

[11] X. Liu, Z. Li, N. Zhao, W. Meng, G. Gui, Y. Chen and F. Adachi, "Transceiver Design and Multihop D2D for UAV IoT Coverage in Disasters", Internet of Things Journal IEEE, vol. 6, no. 2, pp. 1803-1815, 2019.

[12] Fatima Haouari, Ranim Faraj, Jihad M. AlJa'am, "Fog Computing Potentials Applications and Challenges", Computer and Applications (ICCA), pp. 399-406, 2018.

[13] A. Niruntasukrat et al., ``Authorization Mechanism for MQTT-based Internet of Things", IEEE ICC Wksps., pp. 290–95, July 2016.

[14] A. Schmitt, F.Carlier and V.Renault, ``Dynamic bridge generation for IoT data exchange via the MQTT protocol", Procedia Computer Science, vol. 130, pp. 90-97, 2018.

[15] JA. -A. A. Boulogeorgos et al., ``Terahertz technologies to deliver optical network quality of experience in wireless systems beyond 5G", IEEE Commun. Mag., vol. 56, no. 6, pp. 144–151, June 2018.

[16] Zan Li, Lei Guan, Chenxi Li and Ayman Radwan, "A Secure Intelligent Spectrum Control Strategy for Future THz Mobile Heterogeneous Networks", IEEE Communications Magazine, vol.56, no.6 pp. 116 - 123, 2018.

[17] E. Bj örnson, L. Sanguinetti, H. Wymeersch, J. Hoydis, T. L. Marzetta, ``Massive MIMO is a reality—What is next? Five promising research directions for antenna arrays", 2019. [Online]. Available: https://arxiv .org/abs/1902.07678

[18] C. Jiang, H. Zhang, Y. Ren, Z. Han, K.-C. Chen, and L. Hanzo, ``Machine learning paradigms for next-generation wireless networks", IEEE Wireless Commun., vol. 24, no. 2, pp. 98–105, Apr. 2017.

[19] R. Li et al., ``Intelligent 5G: When cellular networks meet artificial intelligence", IEEE Wireless Commun., vol.24, no.5, pp.175–183, Oct.2017.

[20] Q. Mao, F. Hu, and Q. Hao, ``Deep learning for intelligent wireless networks: A comprehensive survey", IEEE Commun. Surveys Tut., vol. 20, no. 4, pp. 2595–2621, 2018.

[21] Ryan C. La Brie and Gerhard H. Steinke, ``Towards a Framework for Ethical Audits of AI Algorithms", Twenty-fifth Americas Conference on Information Systems, 2019. [Online]. Available: https://aisel.aisnet

ISBN 978-981-14-4787-7
Proceedings of 2020 the 10th International Workshop on Computer Science and Engineering (WCSE 2020)
Yangon (Rangoon), Myanmar (Burma), February 26- February 28, 2020, pp. 130-135

A Study On Deep Learning Based Real Time Road information Monitoring Device for Emergency Vehicle Guidance

Seona Park[+], Junghan Ha, Muwook Pyeon, Wonwoo Jung

Konkuk University, Republic of Korea

Abstract: This study is to secure the golden time for emergency vehicles, especially to shorten time in narrow roads. These days, in Korea the black box install rate is over 88.9% and it shows that Koreans use black box much more than other countries. Through this study we can analyze the road situation by collecting the data through the black box and transmit it to the server. And not just guiding the emergency vehicle to the shortcut but to the optimal route when they pass through alleys. Finally we could suggest the solution for the emergency vehicle to be in golden time. For this study we used Raspberry-Pi and Yolo v3 algorithm. There were difference between analyzing the image and the video and concluded that the video take much more time due to the transmission delay and data size.

Keywords: Deep Learning, Emergency Vehicle Guidance, Real Time Road Information, Yolo v3

1. Introduction

It is important for the fire truck to arrive on the scene in golden time(5min) [1]. It has a big difference between the arrival of the fire truck in golden time and over the golden time. The casualties increase of 1.48 times per one fire accident and amount of damage increase of 3.63times if the fire truck doesn't arrive in golden time [2].

In the case of Korea the on-time arrival rate of fire trucks is only 59.3%. And the on-time arrival rate of Jeonbuk area in Korea is 5% less then the average rate and ranked very low above all the districts in Korea. These are due to shortage spatial information and lots of narrow roads in these areas [3].

So we targeted this area and progressed the research.

We've realized that vehicle black box install rate was over 88.9% which shows the high rate of black box usage [4]. So to find the solution for this problem, we used the black box which almost people use, and collect the video by the black box and send it to the server. Then we used the Yolo-v3 algorithm to see if the device could recognize the object and perceive the alley situation [5].

Through this study, we expect the emergency vehicle to be in golden time and minimize the damage. And also this could be helpful for application to the Smart-city system and create a new market that uses the database of spatial big data. Especially we used the mini computer 'Raspberry-pi' and the open source algorithm 'Yolo v3' for image processing. Our goal of this study is shown as Figure 1 below.

2. Trends of the Study

2.1. TCS(Transportation on Control System)

In Eastern part of U.S. offers signal system. TCS uses GPS system to shorten response time. It make use of real-time road information and location of the emergency vehicle so it makes possible to change the signals

[+] Corresponding author. Tel.: + 82-10 85605094
E-mail address: tjsdk7009@gmail.com

green which is in the optimal path to the scene of the accident [6]. The example of TCS is shown as Figure 2 below.

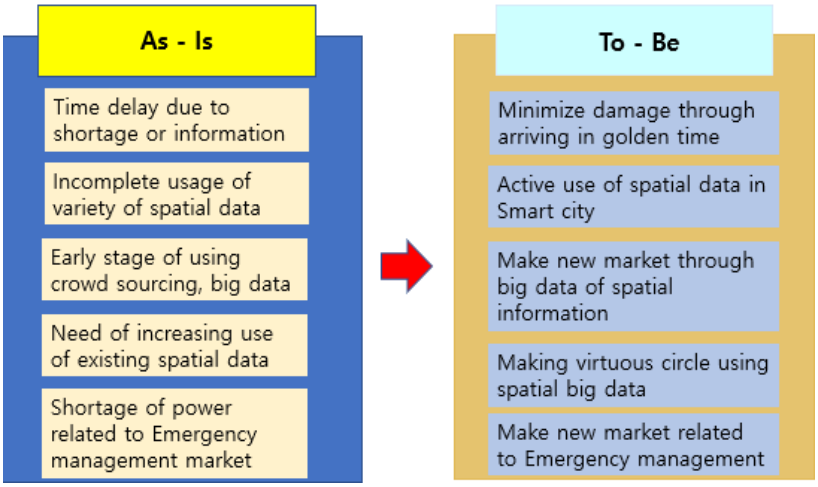


Fig. 1: Goal of this study

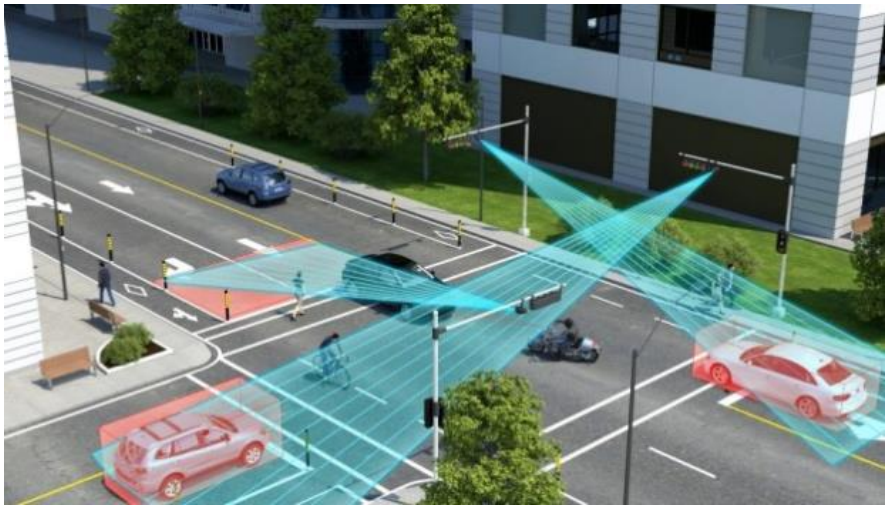


Fig. 2: Transportation on control system (TCS)

2.2. LBS(Location Based Service) Technology used by Lots of Companies

Apple is serving Indoor Map and navigation service and the apple map is leading the evolvement of Apple map to the LBS platform. Figure 3 shows the process of APPLE LBS service.

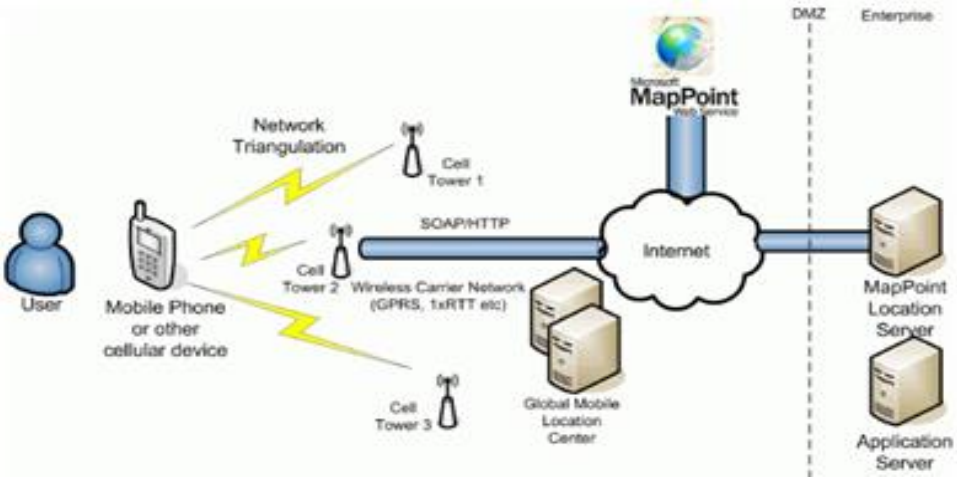


Fig. 3: Composition of APPLE LBS

Google is serving location service by google map.

HERE is serving LBS map for vehicle, Enterprise and working hard to develop related technology in self-driving car with LG and SKT [7].

3. Road Information Collecting Device

3.1. The Device We Used in the Experiment

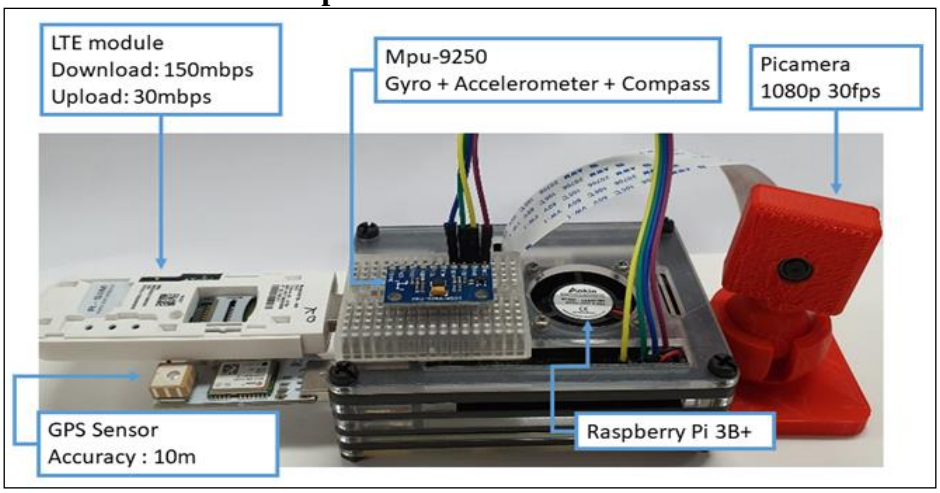


Fig. 4: Black box device used in the experiment

We made the device to have the similar specification of the black box shown as Figure 4. The black box need processor, GPS, position recognition system, camera and additionally we used communication device to communicate with the server. For the whole control, we used raspberry-pi 3B+ [8]. We chose the Raspberry pi because it is more frequently used then other developer devices and we thought it had a sufficient specification to make the black box.

For the camera, we used the Pi-camera that could fit in to the raspberry-pi. Pi camera offers maximum 8MP 3280*2464 Pixel, but these days lot of black box uses 4MP 1920*1080 Pixel, so we lowered the specification of the Pi-camera to fit in to the general black box specification.

Other devices and models are composed as the table 1 below [9].

Table 1: Information of The Device Used In The Experiment

Composition	function	performance
Raspberry Pi 3B+	Black box, image processing, communication	CPU : Quad 1.4GHz RAM : 1GB SDDRAM GPIO : 40
LTE module	LTE network communication	Communication velocity : 150mbps LTE communication band : Band1(2100MHz)/Band3(1800MHz)/Band28(700MHz) 3G communication band : UMTS B1/B5 UMTS (2100/850)MHz
GPS Sensor Neo-06	Identify the location of the vehicle, GPS	Accuracy : 1.8m Search channel : 66 Trace channel : 22 Update cycle : 10hz
Mpu-9250	Identify the position of the vehicle through 9 axis	Voltage 3.3v – 5v Communication method : I2C, SPI Gyro : ±250 500 1000 2000/s Acceleration : ±2 ±4 ±8 ±16g Magnetic : 4800uT

Pi camera	Black box camera module	Resolution : 1080p 30fps 1920 * 1080 pixsel Image quality : 8million pixel Viewing angle: -20 °-60 °
Micro SD	Storage	Storage space : 128GB Read : 100MB/s Write : 100MB/s

3.2. Operation Process

Raspberry Pi transmits the data to the server every minute.

The data collected by each devices is composed of (value of GPS(x,y), Pi Camera(pnag, h.264(image extension)), MPU-9250(acceleration, gyro, terrestrial magnetism), Raspberry-Pi network(IP), image(photo,video)) [10]. The Server organize the divided text of the data transmitted from the raspberry pi separately. And analyze the video through the yolo algorithm then store the result. The whole system is shown as Figure 5 below.

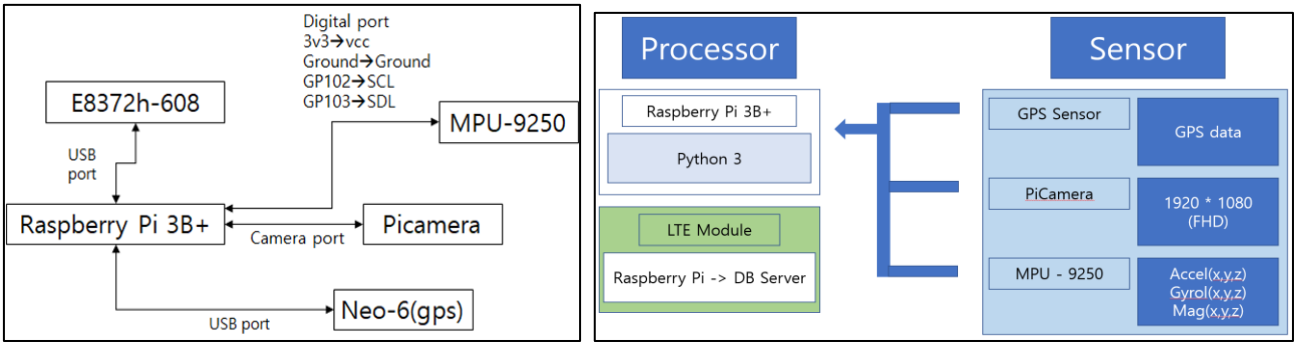


Fig. 5: System configuration

4. Vehicle Feature Extraction by Deep Learning

Fig. 6: Extracting vehicle from the road

Fig. 7: Route where we collect the traffic data

We've drived over 16.805Km to collect the traffic data. The image about vehicle extraction and route is shown in Figure 6 and 7 and the device was installed as Figure 8.

We used the Pi camera of the Raspberry pi to transmit FHD resolution file. And we conducted the experiment in 2 ways.

1. Recognizing photos of FHD quality per 30 seconds.

2. Transmitting 5 second length of videos of FHD quality per 45 seconds.

We used the communication module in the raspberry pi to transmit photos and videos. There was a delay in processing to collect and transmit video from the raspberry pi due to the limitation of hardware. To solve this we put 45 second of waiting time for 5second video so the raspberry pi could have relax time to finish the process.

In communication, nowadays LTE communication uses private network so at the server to communicate directly with the raspberry pi TCP/IP communication is needed. In this case if the black box is connected with lots of routes, the server communication won't be smooth. So it will be appropriate to use http communication for raspberry-pi to transmit data to the server and the server to check the transmission [11].

Lastly the Raspberry-pi's problem is the micro sd used as storage. There is a limitation of speed for micro sd to read and write the data and it is highly insufficient to overcome the difference of the speed with the CPU. To solve this problem the hardware, especially the memory should be developed.

Fig. 8: Device installed in the vehicle

5. Experimental Result and Perspective

5.1. Result

Through the experiment we could send the video and vehicle data to the server completely by combining the Raspberry-pi and LTE router and was possible to recognize the object by the yolo algorithm as Figure 9 below.

But the video processing took a long time. We think GPU is need to make the process faster.

We expect this could be applied to the emergency vehicle guidance. It can recognize the object and decide whether the vehicle is in the road or not. But it has a limitation of deciding the width of the road and the width of the remaining width excluding the parked vehicles, so we need technology to make those kinds of problems solved.

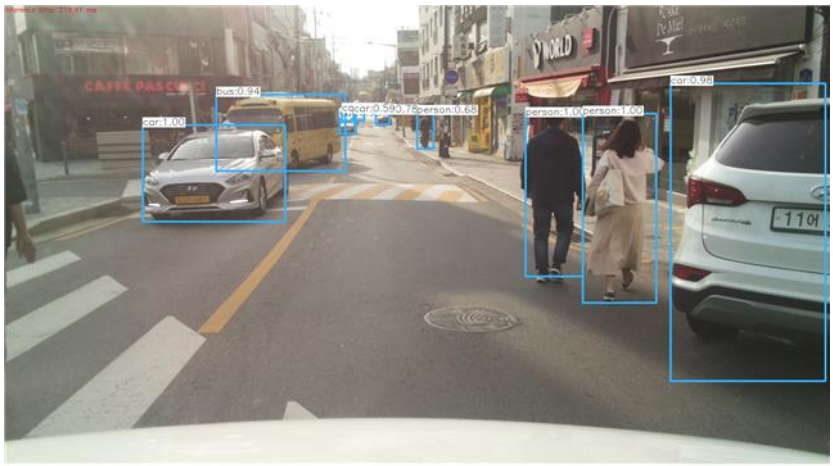


Fig. 9: Object detection through the device

The accuracy of yolo v3 is about 70% but in our study we've got 1~2% lower than that. The reason why is because we've took the video at the front side of the vehicle so it had a different angle of view with the existing data.

5.2. Ways to Improvement

It’s still insufficient to recognize the affordable width the vehicle can pass through. This is because the variety of form of the alleys, different image composition and the location of the black box. We need more precise and standard black box location and develop the algorithm to recognize the road width.

6. Conclusion

We proposed the application of emergency vehicle guidance system using video imagery collected by vehicle black box. It showed a sufficient performance which can distinguish cars and other obstacles clearly. But it still has a limitation of deciding the width of the road and remaining space excluding the parked vehicle.

To solve this issue we need more detail location information and posture of the black box camera and more data of variety of road situation and alleys. Also development of algorithm and GPU are required.

It is meaningful that we can use deep learning in public interests and this may save more lives in emergency situations. GPU and deep learning algorithms are keep being developed so we expect the limitation of this study will be solved within a short time.

7. Acknowledgement

"This research was supported by the MISP(Ministry of Science, ICT & Future Planning), Korea, under the National Program for Excellence in SW(No. 2018-0-00213, Konkuk University) supervised by the IITP(Institute of Information & communications Technology Planing & Evaluation)"(No.2018-0-00213, Konkuk University)

8. References

[1] Choeung Rae. (2015). “Safe Gyeonggi-do: Securing Golden Time.” Issues & Diagnosis, (179), 1-25.

[2] Yun-gu Kang “(A)study on factors restrained from fast moving of fire engine : focused on reduction of moving time” Kangwon National University 2016

[3] Ko Eun-hee '64% 'Arrival within 5 Minutes "Desperate Needs" https://news.joins.com/article/15050078, 2014

[4] Sejin Kim “ Installation of car black boxes increases year by year ... installation rate 89%” http://www.datasom.co.kr/news/articleView.html?idxno=99167. 2019

[5] Shafiee, Mohammad Javad, et al. "Fast YOLO: A fast you only look once system for real-time embedded object detection in video." arXiv preprint arXiv:1709.05943 (2017).

[6] Cheol ki lee , “Emergency vehicle priority signal operation plan”, Police Science institute 70p 2014

[7] Korea Internet & Security Agency “Domestic and foreign LBS industry trend report” 2018

[8] Upton, Eben, and Gareth Halfacree. Raspberry Pi user guide. John Wiley & Sons, 2014.

[9] Pi, Raspberry. "Raspberry pi 3 model b." *Online}. Tillgänglig: https://www. raspberrypi. org/products/raspberry-pi-2-model-b/{Använd 10 02 2016}* (2015).

[10] Yao, Leehter, et al. "An integrated IMU and UWB sensor based indoor positioning system." 2017 International Conference on Indoor Positioning and Indoor Navigation (IPIN). IEEE, 2017.

[11] Tran, Thien-Toan, Yoan Shin, and Oh-Soon Shin. "Overview of enabling technologies for 3GPP LTE-advanced." *EURASIP Journal on Wireless Communications and Networking* 2012.1 (2012): 54

Chapter 5
Advanced Information Technology and Management

ISBN 978-981-14-4787-7
Proceedings of 2020 the 10th International Workshop on Computer Science and Engineering (WCSE 2020)
Yangon (Rangoon), Myanmar (Burma), February 26- February 28, 2020, pp. 136-140

A Study on Community Overlapping Detection Algorithms in Social Networks

Eaint Mon Win [1+], May Aye Khine [2]

[1, 2]University of Computer Studies, Yangon, Myanmar

Abstract. Community detection is one of the most important research area wherein invention and growth of social networks. Community is a set of members densely connected within a group and sparely connected with the other groups. In social networks, the singular characteristic of communities is multi membership of a node resulting in overlapping communities. Another relevant feature of social networks is the possibility to evolve over time. In recent years, many researchers have worked on various methods that can efficiently unveil overlapped structure on dynamic network. This paper reviews the previous studies done on the problem of overlapping community detection algorithms. Moreover, some approaches for dynamic network that change from time to time are also described.

Keywords: overlapping community, community detection, dynamic, social network.

1. Introduction

With the development of information technology, people living in society start to use social media as a virtual environment to find online friends. Social media refers to the interaction of users by exchanging idea in virtual communities. Social Networks are designed as graphs consisting of nodes where each node represents individual user such as people, organizations, and edges or lines between these nodes represents their relationship like friendship or mutual interaction. One of the problems in the study of large complex networks is the detection of community structure i.e. the decomposition of a network into groups (clusters or modules) consisting set of nodes. The process of uncovering these groups is called community detection. It is a clustering approach and difficult task in social network analysis [1].Community detection is used in various applications where group decisions are taken, e.g., delivering information within group or recommending products to group. It can be also used for Target Advertising, Criminology, Public health and Politics [2]. Moreover, community detection or clustering with increasing number of smart devices is applied on smart devices network to generate the device social relation without human intervention by clustering approach of device interactions [3].

Much effort in identifying efficient and effective methods for community detection has been lead on finding disjoint communities. However, communities in real networks may overlap, i.e. a node belongs to one or more communities because an individual (node) share many things in common (e.g. regions, topic, hobbies, and interest). Fig 1 shows the visualization of overlapped community structure with three groups. Detection of overlapped structure becomes a challenge because traditional algorithms that detect only disjoint communities are not suitable. Therefore, overlapping community detection algorithms are defined on static network. In recent year, many researchers analysed overlapping community detection algorithms and evolution of community in social networks [4] [5]. In this paper, a description of the recent developed overlapping community detection algorithm is given and also represents developed algorithms on dynamic networks due to traditional overlapping community detection algorithms fails the problem on dynamic nature of social network.

[+] Corresponding author. Tel.:+(09262639675)
E-mail address: eaintmonwin@ucsy.edu.mm

This paper is organized as follows. The next section describes static and dynamic communities. Section3 considers categorised overlapping community detection methods and proposed methods by many researchers in recent year. Section 4 describes common use network dataset in literatures. Finally, conclusion and future research are represented.

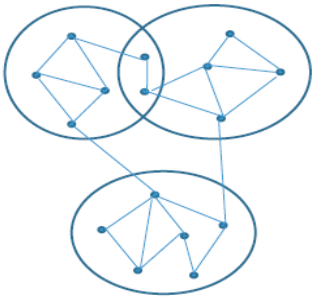


Fig. 1: Visualization of two overlapping communities with three groups

2. Static and Dynamic Communities

Since evolving communities are occurred as dynamic networks, a new research area concerned with dynamic structure has appeared. Communities having dynamic structure can change after a timestamp. For example, let's imagine a social network of American college football, and communities are identified on it. The static community represents all players in community at initial state t0. However as time goes on, players may change such as leave or enter into team and initial players will not be in the team; but, the corresponding community still exists. These dynamic network can be represented by a time-sequence of static networks called time frames (snapshots), each snapshot corresponding to interactions derived from data collection within a time period in Fig 2.

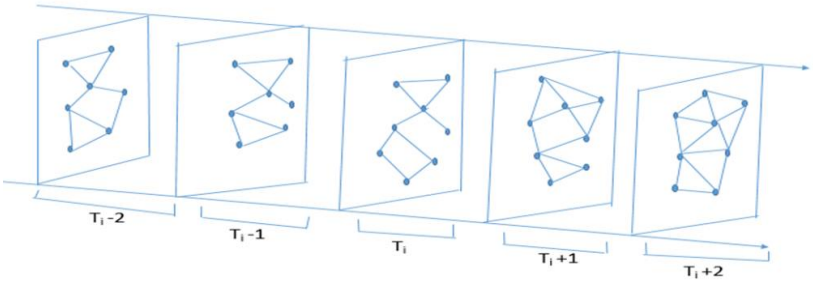


Fig. 2: A temporal network changes within five time frames

2.1. Events of Dynamic Communities

The changes of dynamic network are appearance and disappearance of nodes and edges in network over time. These operations or events are quite simple, but some operations are complex due to changes of entire or partly community. Therefore, implementation on all operation are not easy [6]. These operations on communities are (1) Growth: community's size grows when new nodes come into a community. (2) Contraction: community's size decreases when loses some nodes from community. (3) Merging: one or more communities are merged into one to form a new community. (4) Splitting: a community is decomposed into two or more communities. (5) Birth: a community emerges as first appearance at a given time. (6) Continue: community remain unchanged. (7) Death: a community disappears when loses its all members.

3. Overlapping Community Detection Algorithms

There are many algorithms for detecting community structures. However, most of the traditional algorithms are focused on identifying disjoint communities. Therefore many researchers tend to research the algorithms for detecting overlapping communities. The algorithms are categorized as Clique Percolation, Link Partition, Local Expansion, Label Propagation and Dynamic Networks.

3.1. Clique Percolation

The basic idea of Clique Percolation method to detect overlapping communities is creating a set of many cliques (completely connected subgraph) as communities. Firstly, construct vertices of the k clique graph and then construct edges of the k clique graph (percolate if two vertices in k-clique graph have strong connections). In the new graph, each connect component of the clique graph is a community [7].

Greedy Clique Expansion (GCE) [8] is proposed to detect accurate overlapped communities and get good performance on synthetic data. It begins finding core nodes by detecting maximum cliques and then core nodes are used to obtain community C by inserting node until added node obtain lower fitness. Extended Clique Percolation Method (ECPM) [9] is developed to address problem that cannot discover complete network of CPM. This method finds initial communities and lists left out nodes which are not included in any initial community and calculate their belonging coefficient. Communities are expanded by adding left out into corresponding community according to their belonging coefficient. Finally much similarity communities are merged into single one using Jaccard Similarity. But most algorithms based on CPM is high complexity because it finds many small cliques to get community.

3.2. Link Partition

Link partition method partitions the links in the original graph into several communities. It does not need to know prior knowledge and extracts the link graph using edges instead of nodes. After that graph partition algorithm is applied to cluster or some link community similarity functions are used to directly cluster.

A link clustering based novel algorithm [10] is proposed for loosely connected network that cannot be solved by CPM and for inaccurate overlapping community detection in weak-tie membership. In this paper, algorithm LinkSCAN based on link-space transformation (transforms original graph into a link-space graph). Overlapping Communities are detected using disjoint community detection algorithm on line graph and link similarity on original graph. By using this framework LinkSCAN$*$ that enhances the efficiency of LinkSCAN by sampling. [11] developed a link based clustering algorithm called meme link which optimizes modularity density function to find densely connected links among communities by using weighted graph via similarity function. An efficient algorithm LEPSO [12] is proposed for solving problem of traditional algorithm based PSO that generate superfluous small communities, The algorithm is based on line graph theory, ensemble learning and particle swarm optimization.

3.3. Local Expansion

Local seed expansion method selects seeds and expand selected seed by using various fitness functions and then generates overlapping communities by merging intermediate communities into global communities. The problem to deal with this algorithm is to find good seed sets. In Local Expansion, each implementation differs a lot by depending various fitness function.

Therefore, an algorithm represented in [13] is seed set expansion algorithm to detect overlap communities. The basic concept of algorithm is to identify good seeds, and then expands these seed sets using the personalized PageRank clustering procedure. This algorithm consists of four phases namely filtering (use graph partitioning), seeding (use Graclus Center and Spread Hub), seed set expansion (use PageRank) and propagation to find overlap community. In 2016, [14] detected the minimal cluster using density function to find the nodes that are closely connected with the initial nodes and then finds the local community extended from the minimal cluster. In 2018, proposed an algorithm, OCDNW to find a good seed. The algorithm consists of three parts: initial seed selecting, local community expansion and community optimization. For initial seed selecting, choose seed node with highest weighting of each node by sum score of all edges which connected to the neighbors. Then expand seed to form local community using node fitness function. Finally, merge two communities into larger ones to improve the quality of community if there are too many overlapped vertices between two communities [15].

3.4. Label Propagation

Label propagation algorithm refers to labels propagate between nodes. It begins assigning each vertex to unique label and these labels propagate through the network. In updating labels, ties are broken randomly if the node receive multiple labels. But LPA can detect only disjoint community.

COPRA [16] is a multi-label propagation algorithm to detect overlapping communities. It begins by giving unique label with belonging coefficient setting 1 to every vertex. Each vertex updates its labels by

summing and normalizing the belonging coefficients of vertices in the neighbor set. After propagation, each vertex has multiple labels. Another extension of label propagation technique, SLPA [17] is developed to avoid detecting only disjoint community. It can detect overlapping communities because receives multiple label. In SLPA, each node is initialized unique label and then one node is selected as listener. Each neighbor of the selected node randomly selects a label and sends the selected label to the listener. Listener accepts one of the propagated labels according to listener rule. Finally, threshold is used to generate overlapping communities as post processing. In 2018, [18] developed LPANNI (Label Propagation Algorithm with Neighbor Node Influence) to overcome weakness such as low accuracy, instability (some algorithms based on LPA require to set parameter as priori). It detects overlapping community structures by adopting fixed label propagation sequence based on the ascending order of calculated node importance and label update strategy based on neighbor node influence and historical label preferred strategy (i.e. using idea of COPRA for detecting overlap and DLPA for reducing complexity).

3.5. Dynamic Networks

Most real world social networks are inherently dynamic, growth rapidly in term of social interaction. With community structures change from time to time in evolving network, this research area is receiving more interesting from researchers. The more recent methods aiming to find out dynamic communities are described.

SLPA Dynamic (SLPAD) is an algorithm based on SLPA and incorporates the ability for this algorithm to handle dynamic networks. It involves running SLPA on communities that change from one timestamp to the next. But SLPAD considers only updates based on edge changes not node changes [19]. In 2018, an overlapping community's detection method is proposed using agent that observe the network and consequently update their communities on dynamic networks. For detecting communities each node is iteratively reassigned to the community that yield the highest positive gain by considering similarity attributes in initial partition. It allows all operation on communities: birth, death, growth, and contraction, merge and split [20]. OLCPM is based on clique percolation and label propagation methods. The paper introduced online CPM (OCPM) by building upon CPM to identify the core nodes of communities in real time and it aimed on analysing the dynamic behaviours of the network which may appear from inserting or removing. As post-processing based label propagation, OLCPM is set out on the generated communities of OCPM to discover the peripheral nodes [21].

4. Standard Dataset for Testing Algorithms

The most commonly use dataset in literatures for testing algorithm are divided into two types: real world network dataset and synthetic dataset. Mostly use real world dataset and available sites are shown on Table 1. For Synthetic dataset, suitable dataset for overlapping community detection algorithms in literature is LFR benchmark (an algorithm generate benchmark dataset is proposed by Lancichinetti) [22].

Table 1: Available network dataset for testing algorithms

Type	Network	Number of Node	Number of Edge	Available site
Real World Network	Zachary Karate Club	34	78	http://www-personal.umich.edu/~mejn/netdata/
	Bottlenose Dolphin	62	159	
	College Football Club	115	613	
	US Politics books	105	441	
	Ego Facebook	2888	2981	https://snap.stanford.edu/data/ego-Facebook.html
Synthetic Network	LFR benchmark			https://sites.google.com/site/santofortunato/inthepress2

5. Conclusion

Community detection is more and more attention in social network with the rapidly growth of social data. Till present, related research area is still popular due to challenges of community detection such as dynamic nature, overlapping nature of social network and algorithm instability. This paper reviews algorithms for detection of overlapped communities and algorithms on dynamic networks. As future work, modified

algorithms are developing for instability of algorithm and leads to purpose for defining efficient and effective methods on dynamic network. As data can get large and distribute on web, researchers are comparing accuracy and performance of algorithms on Sparse and Hadoop architecture.

6. References

[1] L. Tang, H. Liu, Community detection and mining in social media, Synthesis lectures on data mining and knowledge discovery . pp:1-137, 2010

[2] A. Karatas, S. Sahin, Application Areas of Community Detection: A Review. 2018 IBIGDELFT

[3] K. Dong-Oh, C. Jang-Ho, J. Joon-Young, and B. Changseok, "Clustering Approach of Device Interactions for Automatic Generation of Device Social Relation," International Journal of Machine Learning and Computing, 5, no. 2, pp. 132-136,2015

[4] J. Xie, S. Kelley, B.K. Szymanski, Overlapping community detection in networks: The state-of-the-art and comparative study, 2013, ACM

[5] N. Dakiche, FBS. Tayeb, Y. Slimani, and K. Benatchba, Tracking community evolution in social networks: A survey Information Processing & Management. pp:1084-1102,2019, Elsevier

[6] G. Rossetti, R. Cazabet, Community discovery in dynamic networks: a survey. ACM Computing Surveys (CSUR), 2018

[7] S. Fortunato, Community detection in graph. pp:75-174, 2010 ,Elsevier

[8] C. Lee, F. Reid, A. McDaid, N. Hurley, Detecting highly overlapping community structure by greedy clique expansion. arXiv preprint arXiv:1002.1827,2010

[9] S. Maity, S K. Rath, Extended Clique percolation method to detect overlapping community structure. International Conference on Advances in Computing, Communications and Informatics (ICACCI). pp: 31-37. 2014.IEEE

[10] S. Lim, S. Ryu, S. Kwon, K. Jung, J-G. Lee, LinkSCAN*: Overlapping community detection using the link-space transformation, 2014 IEEE 30th International Conference on Data Engineering. pp: 292-303. 2014.

[11] M. Li, J. Liu, A link clustering based memetic algorithm for overlapping community detection. Physica A , 2018

[12] F. Huang, X. Li, S. Zhang, J. Zhang, J. Chen, Z. Zhai, Overlapping community detection for multimedia social networks. IEEE Transactions on multimedia. pp: 1881-1893. 2017

[13] JJ. Whang, DF. Gleich, IS. Dhillon, Overlapping community detection using seed set expansion. Proceedings of the 22nd ACM international conference on Information & Knowledge Management. pp: 2099-2108. 2013

[14] Y. Zhou, G. Sun, Y. Xing, R. Zhou, Z. Wang, Local community detection algorithm based on minimal cluster. Applied Computational Intelligence and Soft Computing.2016. Hindawi

[15] X. Chen, J. Li, Overlapping Community Detection by Node-Weighting. ICCDA 2018, DeKalb, IL, USA. © 2018 Association for Computing Machinery

[16] G. Steve, Finding overlapping communities in networks by label propagation. New Journal of Physics12, 2010

[17] J. Xie, B K. Szymanski, X. Liu, Slpa: Uncovering overlapping communities in social networks via a speaker listener interaction dynamic process. IEEE, pp: 344-349. 2011.

[18] M. Lu, Z. Zhang, Z. Qu, Y. Kang, LPANNI: Overlapping community detection using label propagation in large-scale complex networks. IEEE Transactions on Knowledge and Data Engineering 2018

[19] Aston, Nathan; Hertzler, Jacob; Hu, Wei; overlapping community detection in dynamic networks. Journal of Software Engineering and Applications. 2014

[20] A. Mahfoudh, H. Zardi, M A. Haddar, Detection of dynamic and overlapping communities in social networks. Int. J. Appl. Eng. Res. pp: 9109-9122. 2018

[21] S. Boudebza, R. Cazabet, F. Azouaou, O Nouali, OLCPM: An online framework for detecting overlapping communities in dynamic social networks. pp: 36-51. 2018, Elsevier

[22] A. Lancichinetti, S. Fortunato; Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities PHYSICAL REVIEW E80, 016118 , 2009.

ISBN 978-981-14-4787-7
Proceedings of 2020 the 10th International Workshop on Computer Science and Engineering (WCSE 2020)
Yangon (Rangoon), Myanmar (Burma), February 26- February 28, 2020, pp. 141-145

Ensemble Learning Method for Enhancing Healthcare Classification

Pau Suan Mung [1+] and Sabai Phyu [2]

[1, 2] University of Computer Studies, Yangon, Myanmar

Abstract. Ensemble learning technique is proposed in this paper for better efficiency of healthcare classification and prediction. Healthcare industry is an ever-increasing rise in the number of doctors, patients, medicines and medical records. Medical history records are beneficial for not only individual but also human society. Three popular machine learning algorithms, namely Naïve Bayes, Support Vector Machine and Decision Tree are applied on this history data as base learners. Two forms of ensemble learning namely bagging and boosting are applied with each base learner for better accuracy than using individually. Comparison results are presented and the experiments show that ensemble classifiers perform better than the base classifier alone. Cervical cancer dataset is used as case study.

Keywords: Ensemble learning, Base learners, Machine learning, Bagging and boosting

1. Introduction

Healthcare data is collection of data about patients, doctors, medicines, treatments and history record of patients and these are so large, distributed, complex and grow so fast. Therefore healthcare industry is highly data intensive. Maintaining and analyzing such large amount of data is a big problem. Traditional database management system are inadequate. Therefore healthcare researchers explore some novel approaches for data aggregation and analysis because of the increasing availability large amount of data of healthcare industry. Cervical cancer used as case study of this paper is the fourth most frequent cancer in women. In 2018, new 570,000 cases are infected and that is 6.6% of all female cancer rate. Approximately 90% of this type of cancer deaths are from the countries with income low or middle. The rate of this cancer can be reduced by prevention, early diagnosis. Effective screening and treatment schemes can be used to reduce this rate. Vaccines are also be provided to protect of human papilloma virus that are common cancer-causing types and therefore the risk of this cancer can be reduced [1].

Data mining or knowledge discovery in database is a computational process that can extract interesting patterns in large volume of history data. The main tasks of data mining include association, classification and clustering. Classification is a process of generating a model to predict appropriate classes of unknown data. Two basic classification processes are model construction and prediction [2]. Some classification techniques are emerged in business. Some are Naïve Bayesian, decision tree, regression, k-nearest neighbor and neural network. Classification methods can be used in many areas of business such as medical, economy and industry applications.

Most machine learning techniques have their own specific results. Each of them has its own pros and cons. To use the concern of individual algorithms or to get benefits from many algorithms, ensemble learning become more important because they combine the predictions or results of several machine learning models to get the overall result of a system more efficient than single algorithms. Ensemble methods become popular for their prediction capacity than individual algorithms. They have already proven successful in both unsupervised and supervised learning. In this study, three machine learning algorithms are used as base learners and each of these are combined with boosting and bagging to improve their accuracies.

[+] Corresponding author. Tel.: +959428121862
E-mail address: pausuanmung@ucsy.edu.mm

This paper is presented as follows: related research works are presented in section 2 and theory background used in this system is presented in section 3, in which some machine learning approaches used as base learners are included, namely Naïve Bayes, Support Vector Machines and Decision Tree classification, and then ensemble learning methods, boosting and bagging, are presented. Section 4 is the experimental result and analysis. This paper is concluded at section 5 and further extension is discussed in this section.

2. Related Works

Many machine learning algorithms have been applied in real applications. In healthcare environment, many machine learning algorithms were used in classification of different diseases. To get more accurate prediction or classification, researchers proposed ensemble learning techniques in which more than one algorithm are used.

One of the ensemble learning techniques was proposed in paper [3]. It named EC3 – Combining Clustering and Classification for Ensemble Learning. In this paper, step by step processing of this novel algorithm was presented. Classification and clustering have been successful individually but they had their own advantages and limitations. The author proposed systematic utilization of both of these types of algorithms together to get better prediction results. Its proposed algorithm can also handle imbalanced datasets. 13 UCI datasets for machine learning repository were used and 60% was for training, 20% for testing and other 20% for validation. Six algorithms were used as base classifiers namely Decision Tree, K-nearest neighbours, SVM, Naïve Bayes, Logistic Regression and Stochastic Gradient Descent Classifier. Base clustering methods were DBSCAN, Affinity, Hierarchical, K-Means and MeanShift.

In the paper [2], the authors proposed efficiency and reliability classification approach for diabetes. The real data was collected from Sawanpracharak Regional Hospital, Thailand and this data was analysed with gain-ratio feature selection. Naïve Bayesian, K-nearest neighbours and decision tree classification were used as base learners on the selected features. To apply the ensemble techniques, bagging and boosting were combined on each of these algorithms. Comparison of results of base learners and ensemble learnings were presented. Then the results of each ensemble learning with respective base learner were collected and compared to find the best method for its research work.

Authors of the paper [4] proposed ensemble learning methods to enhance performance of network intrusion detection system and to reduce false positive rate using bagging, boosting and stacking. It proposed a prototype model using some base classifiers combining with ensemble learning methods. Four base classifiers: Naïve Bayes, decision tree, rule induction and nearest neighbours are used in this model. It proved the accuracy of 99% for detecting known intrusion and stacking can reduce the false positive rate with a high significantly amount of 46.84%.

Classification and regression tree (CART) with resampling techniques was used for classifying imbalanced datasets in the paper [5]. The authors introduced a simplified method for learning such techniques. Based on many metrices such as precision and classification on minority and majority data respectively, the proposed method was compared with other methods. Matthews Correlation Coefficient (MCC) was used because it is suitable with imbalanced data and classification metrices were true positive, true negative, false positive and false negative.

To find the hidden knowledge in medical field, a paper [6] proposed the graph based association rule mining and its proposed system was intended to use with large medical database. It included process of data mining such as data warehousing, data query and cleaning, and data analysis. 6549 obstetrical patients records were collected for exploratory factor analysis.

3. Ensemble Learning

Ensemble learning is one of machine learning techniques, in which multiple learning algorithms are used to get better predictive performance than any of machine learning algorithm alone. Ensemble methods are able to obtain more accuracy than the individual classifiers which make them up. The idea of ensemble learning model is that many weak learners are used together to build a strong learner because the

combination of these weak learners generates increased accuracy than each weak learner. It is also known as committee-based learning because multiple classifiers learn to solve the same problem. An ensemble contains many hypothesis or learners which are usually derived from training data by using each of base learning algorithms. Ensemble methods are well known for their ability to boost weak learners [2].

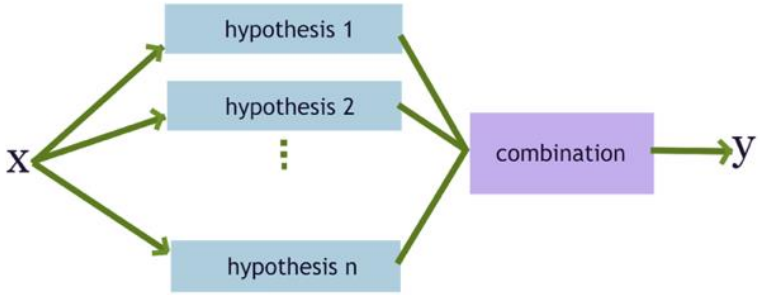


Fig. 1: Ensemble Learning

Two ensemble learning techniques: boosting and bagging, are applied in combination with the base learners. These ensemble learning methods are used to get the results more accurate than using single base classifier.

3.1. Base Classifiers

Three base classifiers: Naïve Bayes, Support Vector Machine and Decision Tree are used in this paper. With these base classifiers, ensemble learning will be applied.

3.1.1. Naïve Bayes

Naïve Bayes classifier based on Bayes theorem is a probabilistic classifier. This classifier is a simple method for building classification model. This classifier is highly scalable and requires a number of parameters (features/predictors). Class labels are assigned to problem instances and the class labels are drawn from some finite set. This classifiers can be trained in a supervised learning efficiently. It can be used in many complex situations in real-world environment. Naïve Bayes classifier is used in application with automatic medical diagnosis [2]. The advantage of this classifiers is that small number of training data is required to estimate for classification. The process of Bayes theorem is mathematical and to find the probability for a condition, that is mostly related with a condition already taken.

3.1.2. Support Vector Machine

Support Vector Machine, SVM, can also be used for analysis of classification and regression and they are supervised learning models in machine learning. Each sample in training data is marked as belonging to one or the other of two categories. This algorithm generates a model to assign new example to one category or the other. This model is a representation of the example as points in space, mapped so that examples of the separate categories are divided by a clear gap that is as wide as possible. Next examples are mapped into the same space and predicted to belong to a category based on the side of the gap on which they fall [7].

3.1.3. Decision Tree

Decision tree can be used for classification and also regression, and they are in the form of tree structure. The data set is divided into smaller and smaller subsets until its leaves arrived and therefore an associated decision tree is an association that is developed incrementally. Decision tree produces decision nodes and leaf nodes at its final result. Each decision node has two or more child nodes or branches. The leaf node is a decision or final result. The topmost decision node is root node. The root node is best predictor. A decision tree is also a top-down structure and the topmost is the root node. The data is partitioned into subsets that have similar values (homogenous). Entropy value is used in decision tree algorithm to get the homogeneity of a subset. The entropy is zero for the sample with completely homogeneous and entropy value equals one for the sample divided equally [8].

3.2. Boosting

Boosting is primarily used to reduce the variance and bias in a supervised learning technique. The idea of this method is to build the weak classifiers repeatedly to be correlated with the true classification. This technique can reduce the error caused by weak classifier significantly. It refers to algorithm family that converts weak learners (base learners) to a strong learner. Although the potential results are estimated by theoretical perspectives, the true value can only be obtained by applying the technique in the real world classification problems [9].

3.3. Bagging

Bagging or Bootstrap Aggregating can be used to improve the accuracy and makes the model more generalize by reducing the variance i.e. by avoiding overfitting. In this, multiple subsets of training dataset are taken and each subset is used to build the classifier. Then the outputs are combined by using voting to get the final decision with more accuracy. Variance can be reduced and overfitting can be avoided. Although this method is applied in decision tree normally, it can be used with other algorithm as well [9].

4. Experimental Result and Analysis

Ensemble learning proposed in this paper use three machine learning algorithms: Naïve Bayes, Support Vector Machine (SVM) and Decision Tree as base learners, and two ensemble methods: Boosting and Bagging are used on each base learner. Firstly the accuracies of these base learners are recorded and then boosting is applied on the base learners, and finally bagging is applied on each of them. At each step, the accuracies are recorded and the comparison of these accuracies is shown in Figure 2.

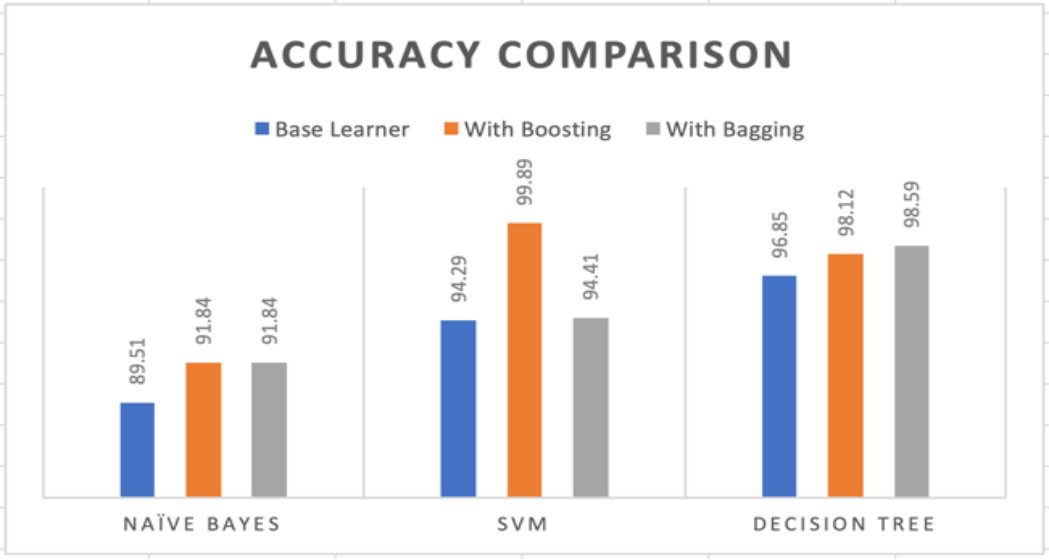


Fig. 2: Accuracy Comparison

This experiments are done with Weka Opening Source Machine Learning Tool running on Java language and the dataset 'Cervical cancer' is taken from UCI machine learning repository [10].

This experiment shows that Naïve Bayes produces least accuracy than other two methods. Although the accuracy of SVM base learner is less than those of Decision Tree, Boosting with SVM produces the most accuracy than other methods. For Bagging method, Decision Tree has the most accuracy than those of other two methods. In all methods, ensemble learning, combining with boosting and bagging, has more accuracy than base learner alone.

5. Conclusion and Further Extension

This research work evaluate the accuracies of various classification models. Cervical cancer dataset obtained from UCI is used as case study. Three base learners or classifiers: Naïve Bayes, SVM and Decision Tree, and two ensemble methods: Boosting and Bagging are used in this study. The result of this experiment revealed that SVM produces best accuracy for Boosting and Decision Tree produces best accuracy for

Bagging. This experiment also shows that the ensemble methods get better performance than its base learners alone. This finding of experiment are useful for choosing the classification algorithm for future application and ensemble learning can also be used for better accuracy in application. Other classification algorithm, stacking, can be used and it is further aspect of this work.

6. References

[1] World Health Organization - WHO, Cervical Cancer, https://www.who.int/cancer/prevention/diagnosis-screening/cervical-cancer/en

[2] Nongyao Nai-arun and Punnee Sittidech. Ensemble Learning Model for Diabetes Classification, *Trans Tech Publications, Switzerland,*Advanced Materials Research Vols. 931-932 (2014) pp. 1427-1431

[3] Tanmoy Chakraborty. EC3: Combining Clustering and Classification for Ensemble Learning, *IEEE International Conference on Data Mining,* 2374-8486/17, IEEE 2017

[4] Iwan Syarif, Ed Zaluska, Adam Prugel-Bennett and Gary Wills. Application of Bagging, Boosting and Stacking to Intrusion Detection, *School of Electronics and Computer Science, UK and Electronics Engineering Polytechnics Institute of Surabaya, Indonesia.*

[5] Supajittree Boonamnuay, Nittaya Kerdprasop and Kittisak Kerdprasop. Classification and Regression Tree with Resampling for Classifying Imbalanced Data, *International Journal of Machine Learning and Computing,* Vol. 8, No. 4, pp. 336-340, 2018

[6] Wael Ahmad AlZoubi, Mining Medical Databases Using Graph based Association Rules, *International Journal of Machine Learning and Computing,* Vol. 3, No. 3, pp. 294-296, 2013

[7] Rohith Gandhi, Support Vector Machine – Introduction to Machine Learning Algorithms, 2018, http://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

[8] Rutgers, School of Art and Science, Decision Tree Regression, https://www.saedsayad.com/decision_tree_reg.htm

[9] Aporras, What is the Difference Between Bagging and Boosting, 2016, https://quantdare.com/what-is-the-difference-between-bagging-and-boosting/.

[10] UCI Machine Learning Repository, https://archive.ics.uci.edu/ml/index.php

ISBN 978-981-14-4787-7
Proceedings of 2020 the 10th International Workshop on Computer Science and Engineering (WCSE 2020)
Yangon (Rangoon), Myanmar (Burma), February 26- February 28, 2020, pp. 146-150

A CSP-based Approach to Design a Subnet Solving a Network Construction Exercise for Beginners

Yuichiro Tateiwa[1+] and Yoshifumi Hisanaga[1]

[1]Nagoya Institute of Technology, Japan

Abstract. When creating exercise problems for network construction, teachers confirm whether the problems are solvable, and then they create a correct answer. However, the work is troublesome and may result in some mistakes such as creating unsolvable problems and incorrect answers. This paper proposes a simulator that computes communications in exercise problems, and CSP formulas for solving exercise problems by analyzing the simulator with symbolic execution. In experiments, In experiments, solving the formulas with CSP solver Z3 found correct answers and unsolvable problems.

Keywords: Zeroconf, CSP, SMT, Z3, network construction, e-learning

1. Introduction

It is important to increase the number of network engineers who administer computer networks. These network engineers can develop the infrastructure to develop a ubiquitous network society and to provide new services for it. The experience of constructing a basic network is useful not only to the network administrators but also to the network application programmers and the network system designers.

When creating exercise problems for network construction, teachers confirm whether the problems are solvable, and then they create a correct answer. However, the work is troublesome and may result in some mistakes such as creating unsolvable problems and incorrect answers.

Let us consider a tool that generates network settings sequentially and executes a network simulator with each of the settings. If the tool confirms that one of the network behaviors satisfies all the requirements of the exercise problem, it outputs the setting as one of the correct answers. However, finding a correct answer requires a significant amount of time because there is a large search space that includes combinations of setting values such as Internet Protocol addresses (IP addresses) and subnet masks.

Symbolic execution [1] is an execution method that executes a program with symbolic values as inputs rather than concrete values. Interpreting a program by propagating symbols can analyze the relationships between the input variables and the inner variables, along with the relationships between the input variables and the branch conditions for an execution path. After finding inner variables and an execution path for satisfying all the requirements of an exercise problem in the simulator, the relationships between them and input variables that store a network setting can be analyzed. The relationships are helpful to reduce the search space because of meaning constraints of network settings.

A Constraint Satisfaction Problem (CSP) is formulated by a set of variables, domain of the variables, and constraints of the variables. If CSP solvers find a set of concrete values that satisfy the formulas, the solvers output the set; otherwise, the solvers notify there are no values satisfying the formulas.

This paper proposes a CSP-based approach to find an example of correct network settings. This method consists of the following two parts:

[+] Corresponding author. Tel.: +81527355450
E-mail address: tateiwa@nitech.ac.jp

1) A simulator to calculate communications that are specified in requirements of the exercise problems

2) CSP formulas consisting of network settings as the set of variables, original rules of each parameter as the domains, and the relationships based on the symbolic execution as the constraints

2. Related Work

Zeroconf [2] is a set of technologies that assign an IP address to a new device that is connected to a network. Zeroconf also resolves host names to IP addresses and detects network services in the network. IPv4 Link-local addressing [3] is a technology used in Zeroconf for IP addressing. A host with the technology can search for an unused IPv4 address (169.254/16) in a subnet including itself. The Dynamic Host Configuration Protocol (DHCP) [4] searches for an IP address that is unused in a network that is connected to the DHCP server and is included in the range of the server settings. However, the exercise problems also require design of Media Access Control address (MAC address) assignments and cable connections, which are not supported by the above technologies. The technologies cannot determine whether there is a solution to the exercise problems.

3. Proposed Approach

3.1. Preliminaries

The function *Value(s, e)* in a step *s* and an expression *e* returns the output of *e* just before starting *s*. The function *Value(s, v)* in a step *s* and a variable *v* returns the value of *v* just before starting *s*. The sequence element X_i returns the *i*-th element in *X*. The operator *.(dot)* accesses the value of a member variable in a tuple.

3.2. Network Settings

Table 1 denotes the network setting items and their data structure used in the exercises. The students can set values just to the underlined variables. Hosts equip one Ethernet port named ep1 and switching hubs equip five Ethernet ports named ep1-ep5. It is assumed that communication data can be exchanged between a host and a switching hub.

3.3. Exercise Problems

Exercise problems consist of **communication examples** and a **setting requirement**.

A communication example is an example of Internet Control Message Protocol (ICMP) echo communications, which should be established in correct networks. Let us consider that a device *src* sends an ICMP echo request packet with a destination IP address *dip*. The data structure of a **send setting** is a tuple *(src, dip)*. The data structure of a communication example is a 3-tuple *(snd, OW, HW)* for which *OW* and *HW* are a sequence of devices which are sorted in arrival order of the ICMP echo request and its reply on send setting *snd*, respectively. *OW* and *HW* can be referred to as communication routes.

A setting requirement consists of devices that must be installed into networks, values that must be set in devices, and Ethernet cables that must be connected to the assigned devices, and have the same data structure as that of items in Table 1. All the items must be assigned concrete values, unless underlined items are assigned "*" (denoting arbitrary values), to which students assign concrete values in exercises.

In actual exercise problems, communication examples and a setting requirement are expressed using natural language and figures. Fig. 1 shows an example of exercise problems based on Fig. 2. The upper terms in the squares denote the device type and the lower terms represent the device identifier.

3.4. Communication Simulator

This study proposes a communication simulator for computing communication routes of ICMP echo in Fig. 3 and Table 2. The function *Out* works for transmission of ICMP echo data in a host. The function *Fwd* implements reception and transmission of the data in a switching hub. The function *In* realizes reception and response of the data in a host. The parameters *nd*, *ep*, *pl*, *dip*, *dm*, *sip*, and *sm* of the functions denote a device identifier, an Ethernet port name, a payload, a destination IP address, a destination MAC address, a source IP

address, and a source MAC address, respectively. When establishing ICMP echo communication that satisfies *OW* and *HW*, statements in the simulator are executed as follows:

1. When OW_1 sends an ICMP echo Request *req*, steps 1-8 and 37 are executed. *Value(8, dev2)* is equal to OW_2.
2. When OW_2 receives and sends *req*, steps 9-11 and 36 are executed. *Value(11, dev2)* is equal to OW_3.
3. When OW_3 receives *req*, steps 12-16 and 35 are executed.
4. When HW_1 (i.e., OW_3) sends the ICMP echo reply *rep* (i.e., the response to *req*), steps 17-24, 34, and 35 are executed. *Value(24, dev2)* is equal to HW_2.
5. When HW_2 receives and sends *rep*, steps 25-27 and 33 are executed. *Value(27, dev2)* is equal to HW_3.
6. When HW_3 receives *rep*, steps 28-32 are executed.

Table 1: Network Setting Items and The Corresponding Data Structure.

Name		**Data structure**
Network		Tuple (Set of devices *Devices*, Set of cables *Cables*)
Device	Host	Tuple (identifier *id*, IP address *ip*, subnet mask *mask*, MAC address *mac*, routing table entries *re1*, *re2*, *re3*, ARP cache entries *ae1*, *ae2*, *ae3*)
	Switching hub	identifier *id*
Cable		Set {(device identifier *dev*, Ethernet port name *ep*), (device identifier *dev*, Ethernet port name *ep*)}
Routing table entry		Tuple (destination IP address *ip*, destination network address *nwaddr*, next hop IP address *nh*, *nwaddr*'s subnet mask *mask*, sender Ethernet port name *ep*)
Arp cache entry		Tuple (IP address *ip*, MAC address *mac*)

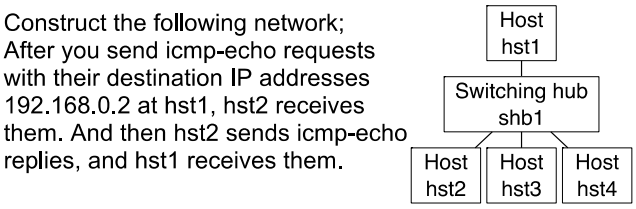


Fig. 1: Example of an Exercise Problem.

```
SR = ({hst1, hst2, hst3, shb1}, {cbl1, cbl2, cbl3, cbl4})
CE = {(snd, OW, HW)}
     snd = (hst1, 192.168.0.2)      OW = <hst1, shb1, hst2>
hst1 = (hst1, *, *, *, re1, re2, re3, ae1, ae2, ae3)
     re1 = (*, *, *, *, *)          ae1 = (*, *)  shb1 = shb1
cbl1 = (hst1, *, shb1, *)
```

Fig. 2: Part of Components of The Communication Example.

Code	Step No.			
def Out(dev, pl, dip)				
(ip, ep) = L3Rtng(dev, dip)	1	17		
if(ep == 'ep1')	2	18		
sip = SIP(dev, ep)	3	19		
if(sip != '')	4	20		
sm = SMac(dev, ep)	5	21		
dm = DMac(dev, ep, ip)	6	22		
if(dm != '')	7	23		
(dev2, ep2) = Next(dev, ep)	8	24		
Fwd(dev2, ep2, pl, dip, dm, sip, sm)			34	37
end				
end				
end				
end				
def Fwd(dev, ep, pl, dip, dm, sip, sm)				
ep2 = L2Rtng(dev, ep, dm)				
if(ep2 != '')	9	25		
(dev2, ep3) = Next(dev, ep2)	10	26		
In(dev2, ep3, pl, dip, dm, sip, sm)	11	27		
end			33	36
end				
def In(dev, ep, pl, dip, dm, sip, sm)				
chkmac = ChkDMac(dev, ep, dm)				
if(chkmac)	12	28		
chkip = ChkDIP(dev, dip)	13	29		
if(chkip)	14	30		
if(pl == 'REQUEST')	15	31		
Out(dev, 'REPLY', sip)	16	32		35
end				
end				
end				
end				

Fig. 3: Communication Simulator

Table 2: Functions Used in The Simulator.

Name	Description
ChkDIP (dev, dip)	If *Obj(dev).ip=dip* is satisfied, this function returns *true*; otherwise, it returns *false*.
ChkDMac (dev, ep, dm)	If *ep='ep1' ∧ dm=mac* is satisfied, this function returns *true*; otherwise, it returns *false*.
DMac (dev, dip, ep)	If *Obj(dev).ae1.ip=dip* is satisfied, this function returns *Obj(dev).ae1.mac*; else, if *Obj(dev).ae2.ip=dip* is satisfied, this function returns *Obj(dev).ae2.mac*; else, if *Obj(dev).ae3.ip=dip* is satisfied, this function returns *Obj(dev).ae3.mac*; else if there is an *ne* for which *Obj(ne.dev).ip=dip ∧ ne.dev≠dev* is satisfied in *ne ∈Neigh(nx.dev)*, *nx=Next(dev, ep)*, this function returns *Obj(ne.dev).mac; otherwise*, this function returns an empty string.
L2Rtng (dev, ep, dm)	If there is an *ne* for which *Obj(ne.dev).mac=dm* is satisfied in *ne ∈Neigh(dev)*, this function returns *Next(ne.dev, ne.ep).ep* for the first found *ne;* otherwise, this function returns an empty string.
L3Rtng (dev, dip)	If there is an *r* for which *r.ip* matches *dip* by the longest prefix match or for which *r.nw* and *r.mask* match *dip* by the longest prefix match in *r ∈{Obj(dev).re1, Obj(dev).re2, Obj(dev).re3}*, this function returns a tuple *(r.nh, r.ep)* (*r.nh≠0.0.0.0* is satisfied), a tuple *(dip, r.ep)* (*r.nh=0.0.0.0* is satisfied), or a tuple *('', '')* (otherwise).
Neigh(dev)	This function returns a set *{nx | nx≠('',''), nx=Next(dev, ep), ep ∈NI(dev)}*.
Next (dev, ep)	If there is a *c* in *{(dev, ep), c} ∈nw.Cables*, this function returns *c*; otherwise, this function returns *('', '')*.
NI(dev)	This function returns a set of Ethernet port names equipped on *dev*. If *dev* is a host, this function returns a set {'ep1'}; else if *dev* is a switching hub, this function returns a set {'ep1', 'ep2', 'ep3', 'ep4', 'ep5'}.
Obj(dev)	If there is a *obj* for which *id=obj.id* in *obj ∈nw.Devices*, this function returns *obj*.
SIP(dev, ep)	If *ep='ep1'* is satisfied, this function returns *Obj(dev).ip*; otherwise, it returns an empty string.
SMac (dev, ep)	If *ep='ep1'* is satisfied, this function returns *Obj(dev).mac*; otherwise, it returns an empty string.

It can be determined whether an answer a*ns* satisfies a communication example *ce* with the following steps:

1. Set up the global variable *nw* of the simulator based on *ans.*
2. Execute the function *Out(ce.snd.src, 'REQUEST', ce.snd.dip)*.
3. If the simulator acts by following the six steps described above, then *ans* satisfies *ce*.

3.5. CSP Formulation

In order to find network settings *nw* that satisfies both setting requirements *sr* and a communication example *ce*, this paper formulates exercise problems as CSPs.

- **Variables:** Empty variables in *nw*, which correspond to the variables whose values are assigned as "*" in *sr*.
- **Domains:** These are shown in Table 3. The variable *dev* in Cable stores an integer identifying hosts and switching hubs, which are assigned with a unique integer that ranges from 1 to the number of devices. The variable *ep* in Cable stores an integer that corresponds to a name of Ethernet port; the integer *i* corresponds to Ethernet port 'ep*i*'.
- **Constraints:** *Eq. 1 ∧ Eq. 2 ∧ Eq. 3* using the following expressions

In order to execute the simulator with *snd*, the following constraints are required.

$$Value(1, nd) = ce.snd.src \wedge Value(1, pl) = 'REQUEST' \wedge Value(1, dip) = ce.snd.dst \quad (1)$$

In order for the simulator to perform steps 1-37, the following constraints are required.

$$\begin{aligned} &Value(2, ep == 'ep1') = Value(4, sip != '') = Value(7, dm != '') = Value(10, ep2 != '') \\ &= Value(13, chkmac) = Value(15, chkip) = Value(16, pl=='REQUEST') = Value(18, ep != '') \\ &= Value(20, sip != '') = Value(23, dm != '') = Value(26, ep2 != '') = Value(29, chkmac) \\ &= Value(31, chkip) = true \wedge Value(32, pl=='REQUEST') = false \end{aligned} \quad (2)$$

In order that the value of variables storing reception devices in the simulator satisfies *ce.OW* and *ce.HW*, the following constraints are required.

$$Value(8, dev2)=ce.OW_2 \wedge Value(11, dev2) = ce.OW_3 \wedge Value(24, dev2) = ce.HW_2 \wedge Value(27, dev2) = ce.HW_3 \quad (3)$$

Table 3: Domains of The Variables.

Variables		Domains
Host	*ip*	[0, 4294967295]
	mask	[0, 4294967295]
	mac	[0, 281474976710655]
Cable	*dev*	[1, size of Devices]
	ep	[1, 5]
Routing table entry	*ip*	[0, 4294967295]
	nwaddr	[0, 4294967295]
	nh	[0, 4294967295]
	mask	[0, 4294967295]
	ep	1
Arp cache entry	*ip*	[0, 4294967295]
	mac	[0, 281474976710655]

4. Experiments

This study implemented a prototype of the CSP formulas with the SMT solver Z3 4.4.1 [5]. The prototype evaluated the formulas with the exercise problem in Fig. 1 and three types of network settings on a PC that featured a 2.93 GHz CPU and a 4 GB main memory.

No. 1 in Table 4 shows the prototype took network settings that had no values and then outputted concrete values that were regarded as correct. No. 2 indicates the prototype took an example of the correct settings and then reported that the settings satisfied the constraints; we agreed with the opinion. No. 3 suggests that the prototype took incorrect settings and then notified that the settings did not satisfy the constraints; we agreed with the opinion. While the prototype ran, the execution times were measured manually. Each execution time was less than one second.

Table 4: Evaluation of The Results.

No.	Type of settings	Output by prototype	Opinion of authors	Execution time (sec.)
1	Empty	Concrete settings	Correct	< 1.0
2	Correct	Satisfactory	Agree	< 1.0
3	Incorrect	Unsatisfactory	Agree	< 1.0

5. Conclusion

A simulator that computes the communications in the exercise problems was proposed in this study. This paper also describes the CSP formulas for solving the exercise problems by analyzing the simulator with symbolic execution. The experiments show that the results for solving the formulas with the CSP solver Z3 are as per our estimations.

In our future work we will focus on expanding this method to deal with larger networks, including routers, and developing a tool for generating CSP formulas based on the exercise problems.

6. References

[1] James C. King, “Symbolic execution and program testing,” Communications of the ACM, Vol. 19, No. 7, pp.385-394, July 1976.

[2] Zero Configuration Networking (Zeroconf), http://www.zeroconf.org, accessed Dec. 16, 2019.

[3] S. Cheshire, B. Aboba, E. Guttman, “Dynamic Configuration of IPv4 Link-Local Addresses,” RFC 3927, May 2005.

[4] R. Droms, “Dynamic Host Configuration Protocol, ” RFC 2131, March 1997.

[5] Z3, https://github.com/Z3Prover/z3, accessed Dec. 20, 2019.

ISBN 978-981-14-4787-7
Proceedings of 2020 the 10th International Workshop on Computer Science and Engineering (WCSE 2020)
Yangon (Rangoon), Myanmar (Burma), February 26- February 28, 2020, pp. 151-155

GPS Trajectory Cleaning For Driving Behaviour Detection System

Tin Lai Lai Mon[+]

University of Computer Studies, Yangon, Myanmar

Abstract. 2017 records of WHO show that road traffic accidents deaths in Myanmar reached 10,527 or 2.67% of total deaths. It also shows that most of the road traffic accidents are due to drivers. Therefore, it is essential to develop a system which is able to detect the drivers' driving behavior to reduce the traffic accidents. Nevertheless, in order to detect the drivers' driving behavior, their accurate GPS trajectory is needed and there are still lots of challenges to get accurate GPS trajectory. GPS trajectory cleaning is one of the most important steps for preprocessing GPS data. GPS data might be wrong because sometimes signals transmitted by satellites cannot be recorded accurately by GPS receivers because of the interference, weak signal, and malfunction of sensor. Therefore, the recorded GPS point can be far from the actual location of the receiver. This inaccurate location can make strong affect to some decision making processes based on the GPS data. In order to eliminate wrong GPS data points from the trajectory, some trajectory cleaning processes: detecting stop points, interpolating missing segments and removing inaccurate points are proposed in this paper. Moreover, the results show that preprocessing trajectory cleaning approach helps to improve the quality of trajectory clustering.

Keywords: Trajectory cleaning, Detecting stop points, Interpolating missing segments, Trajectory clustering.

1. Introduction

Clustering is a processing of grouping similar objects. GPS data can be points or trajectory data. By clustering trajectory data, one can find clusters of objects that follow the same path or detect groups that moved together for a period of time. One can also detect a driver's driving behavior from his trajectory. If it is able to detect a driver's behavior, it is also possible to prevent from road traffic accidents due to drivers by alerting the driver to change his driving behavior at once. But, in order to correctly detect his driving behavior, the accuracy of the trajectory data is crucial. Although there are many methods of trajectory clustering, there is a lack of data preprocessing which is very important to get the result with high accuracy. Critical roles of trajectory data mining in modern intelligent systems are surveillance security, abnormal behavior detection, crowded behavior analysis and traffic control. Clustering algorithms can be categorized into three. They are unsupervised, supervised and semi-supervised algorithms.

In this paper, there will be three steps for cleaning trajectories: stop detection, missing segments and interpolation and inaccuracy removal. Trajectory preprocessing is evaluated by clustering both raw and the cleaned dataset and comparing results. And the results show that the proposed preprocessing gives the better quality of clustering.

2. Proposed System

2.1. Overall System Design

Our proposed overall design of driving behavior detection system is shown in Figure 1. There are 4 stages in our system. The first stage is to collect GPS data, the second stage is to do trajectory cleaning, the

[+] Corresponding author. Tel.: +959-95300903
E-mail address: tinlailaimon@ucsy.edu.mm

third stage is to enhance data by adding weather conditions, road speed limit and important places and the last stage is to measure the aggressiveness of a driver [1]. In this paper, the second stage, trajectory cleaning, will be emphasized.

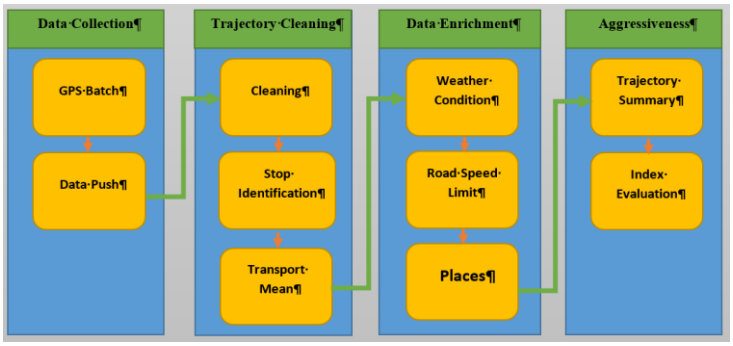


Fig. 1: Overall Design of Driving Behaviour Detection System

2.2. Clustering Methods

Clustering methods can be divided into three main groups. They are model-based clustering; distance-based clustering and visual-aided clustering [2]. In model-based clustering data is considered as coming from a distribution that is mixture of two or more clusters. Each cluster corresponds to a different distribution, and generally, the distributions are assumed to be Gaussians. The parameters of each distribution are estimated by maximizing the likelihood of the expression data. The k-mean approach is a special case of model-based clustering, where all the distributions are assumed to be Gaussians with equal variance. Distance-based methods use distance functions to show similarity between objects. This allows breaking the whole trajectory clustering process into two steps: (1) calculation of distances between trajectories according to the defined distance function and (2) actual clustering using a known clustering algorithm. Finally, visual-aided methods depend on the decision of experts. Experts will change the clustering settings according to their knowledge and experience until the system achieves the desired result [3].

2.2.1. DBSCAN Method

In this work, one of the well-known clustering algorithms, a density-based clustering algorithm called DBSCAN is used to classify every GPS point as CORE, BORDER or NOISE. DBSCAN classification is based on the input parameters: minimum points to cluster (*minPts*) and the radius (*Eps*). At first, *minPts* is necessary to define and every point which has at least *minPts* in radius (*Eps*) can be classified as CORE. A BORDER is a point which does not satisfy the CORE criteria but it is connected to a CORE. Every point which is neither a CORE nor a BORDER is classified as NOISE. An example of a DBSCAN clustering is shown in Figure 2. In this figure, *minPts* =4. Point A and the other red points are core points because the area surrounding these points in the radius (*Eps*) contain at least 4 points (including the point itself). Because they are all reachable from one another, they form a single cluster. Points B and C are also included in the cluster as they are BORDER points. Point N is a NOISE point that is neither a CORE point nor BORDER. This algorithm will be modified in the stop detection step and during actual trajectory clustering.

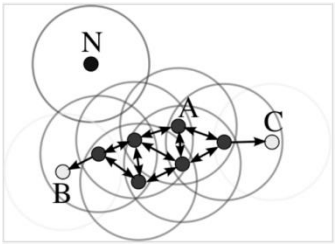


Fig. 2: DBSCAN clustering

2.2.2. Dynamic Time Warping (DTW)

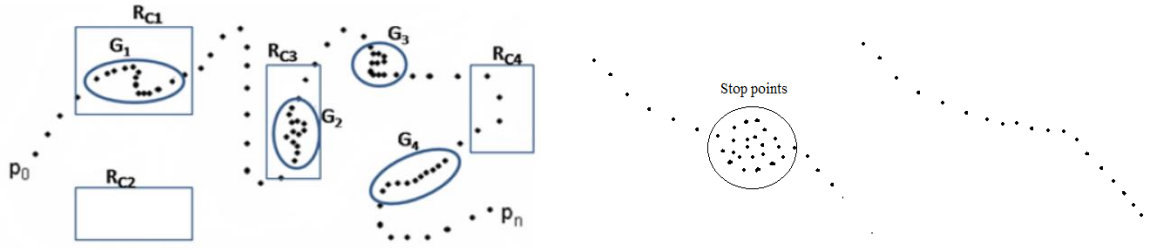
One of the algorithms for measuring similarity between two temporal sequences used for GPS trajectory data is Dynamic Time Warping (DTW). DTW was originally developed for speech recognition. DTW is used to calculate the similarity between objects that have different speeds. Suppose that there are two

sequences of time series. By using DTW, all points of these sequences are warped to each other to minimize the resulting distance. The main advantage of DTW in comparison with Euclidean distance function is that it includes stretching and compression of sequences. In this paper, DTW is used as a distance function that data preprocessing can improve clustering. However, there are also other distance functions to be used.

2.3. Stop Detection

There are many stop detection methods. One of those methods called Intersection-Based Stops and Moves of Trajectories (IB-SMoT) is considered based on the intersection of a trajectory with the user-specified relevant feature types (interesting area or building) for a minimal time duration. Suppose that a trajectory intersects the geometry of an interesting area or building and if the duration of this intersection is greater than or equal to the predefined amount of time, it is considered as a stop point.Another method called Clustering-Based Stops and Moves of Trajectories (CB-SMoT) finds stops nearby places on the trajectory where object spent a relatively large time without leaving those locations. In this work CB-SMoT is applied to find stops and remove them. The purpose of removing these stops is that many distance functions are quite sensitive to them. For example, DTW processes each point in a sequence, the distance between trajectory A (without stops) and similar trajectory B (with stops), would be larger compared to distance between trajectory A and trajectory C (without stops from trajectory B). This may lead to decrease accuracy of trajectory clustering.

In this research work, to detect stop detection three parameters *Eps* which is used in DBSCAN method, *minTime* (time an object spent at particular locations) and area (approximate proportion of points that can form stops) will be selected automatically. In order to detect stop point, it is necessary to find core point condition first. A core point can be defined based on *Eps* and *minTime*, if $minTime \leq T_{last} - T_{first}$, where T_{last} is the latest timestamp and T_{first} is the earliest timestamp in the *Eps*-neighborhood of the core point. The cluster is expanded with all the density-reachable points from the *Eps*-neighborhood. According to the observation, it is found that *Eps* parameter can be easily estimated by taking the mean of all the distances between consecutive points of a trajectory and this mean is sufficient to detect all the major stops on a trajectory. Founded stops are shown in Figure 3(a) where p0,p1,…,pn are GPS points, G1,G2,G3 and G4 are the clusters and RC1, RC2, RC3 and RC4 are candidate stops [4]. After finding all the stops on the trajectory, the stops are moved and filled in the created gap with points calculated based on Missing Segment Interpolation. The comparison of GPS points before and after removing stop points is shown in the following Figure 3(b).



(a) (b)

Fig. 3: (a). Stop detection, (b). Before and after removing stop points

2.4. Missing Segment Interpolation

In trajectory data, there may be some gaps because of the loss of GPS signal. Then, it is necessary to estimate missing segments (missing part of a trajectory according to a given GPS sampling rate and object's movement direction). In this work, missing segments are emulated using a simple interpolation technique.

It is obvious that in order to fill the missing gap it just needs to connect the closest two points before and after the perceived gap. But such kind of consideration leads to the ignorance of the GPS data sampling rate. Moreover, there will be inaccurate results in finding DTW distance.

Suppose P_t and P_{t+1} are consecutive points on a trajectory with timestamps t_{Pt} and t_{Pt+1} and let φ and ψ be the interpolation and trajectory breaking thresholds, respectively. If the time difference between t_{P1} and t_{P+1} is larger than the interpolation threshold φ, this segment is said to be "complete". If this difference is also greater than the second threshold ψ, this segment will not be interpolated. Instead, the segment will be

broken down into two separate trajectories. Moreover, given trajectory will be separated into two sub trajectories if there is no GPS signal for ψ amount of time. This trajectory partitioning is performed even before stop detection. Suppose that

P – the list of points on a trajectory
P_a, P_b – the endpoints of the missing segment and
N – the number of sub segments.

Then, $N = \lceil \frac{2kDist(Pa,Pb)}{\sum_{j=a-k}^{a-1} Dist(P_j,P_{j+1}) + \sum_{j=b}^{b+k-1} Dist(P_j,P_{j+1})} \rceil$

where, $Dist(P_a,P_b)$ is the Euclidean distance between Pa and P_b.

Then, the distance between two consecutive generated points P_i and P_{i+1} (where $a < I < b$) to fill in a missing segment is defined as:

$$Dist(P_i, P_{i+1}) = \frac{Dist(P_a, P_b)}{N}$$

After that, N-1 points are created starting from Pa towards P_b according to the calculated distance $Dist(P_i, P_{i+1})$. After generating all required points, a timestamp for each point is added based on the number of generated segments and on the time difference between two endpoints of the missing segment. In this way, missing segments will be filled using the interpolation algorithm mentioned above. Before and after interpolation of missing segments is shown in Figure 4.

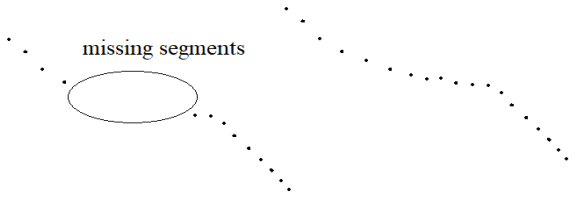


Fig. 4: Before and after doing missing segment interpolations

2.5. Removing inaccurate GPS Points

Some erroneous points exist in many datasets and they are also needed to remove. Erroneous points may be points which have the same coordinates but separated in time. Such kind of erroneous points cannot be found by stop detection if their timestamps is satisfied with *minTime* threshold value. So, the points that have the same coordinates are needed to remove. Some trajectories with null speed and randomly changing their locations with time are also needed to be removed. Moreover, after breaking down the raw dataset into separated trajectories, trajectories which give no meaningful knowledge to the final clustering are also needed to be removed. Therefore, trajectories that did not meet the threshold on the minimum number of points per trajectory are eliminated.

3. Experimental Results

This research proposal is implemented based on the raw trajectory data. From the raw data, sub trajectories are extracted. And then, those sub trajectories are preprocessed using stop detection, missing segment interpolation and inaccurate points removing. And then, there are two datasets: raw dataset and cleaned dataset. Similarity matrices of both datasets are computed using DTW distance between trajectories. Then computed similarity matrices are inputted to DBSCAN. Finally, clustering results of both dataset are compared using a quality measure, QMeasure. To calculate QMeasure the Sum of Squared Error (SSE) and the noise penalty which is to penalize incorrectly classified noises are used. QMeasure is used to minimize the sum of squared pairwise distances between elements that belong to one cluster (Total SSE), while penalizing for incorrectly identified noise points:

$$QMeasure = TotalSSE + NoisePenalty$$
$$= \sum_{i=1}^{|C|} \frac{1}{2|C_i| \sum_{x \in C_i} dtwDist(x,y)^2} + \frac{1}{2|F| \sum_{w \in F} \sum_{z \in F} dtwDist(w,z)^2}$$

where, C is a set of clusters C_i, F is a noise trajectories set and *dtwDist(x,y)* is the DTW distance between trajectories x and y. In this case, *QMeasure* with smaller values means more accurate clustering.

Comparison of *QMeasure* for both dataset of two trajectories is shown in the following Figure 5. In Figure 5(a) and 5(b), Eps value is fixed to 2000 and QMeasure is calculated by changing the minPts from 0 to 10 for Trajectory-1 and Trajectory-2. In Figure 5(c) and 5(d), minPts is fixed to 3 and QMeasure is calculated by changing the Eps values for the Trajectory-1 and Trajectory-2.

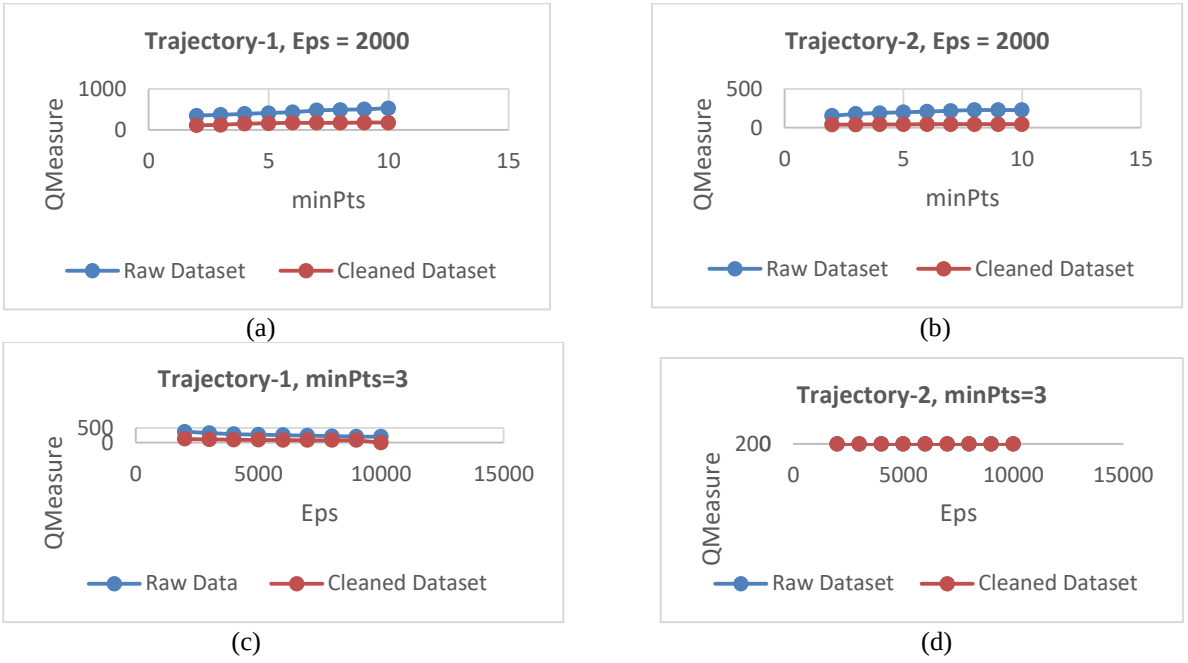


Fig. 5: Comparison of QMeasure between raw dataset and trajectory cleaned dataset

According to the experimental results, it is obviously seen that for any trajectory path with changing *minPts* values and *Eps* values, clustering quality of the proposed cleaned dataset is always better.

4. Conclusion

In this paper, the second stage of the research work, a trajectory cleaning process using trajectory clustering methods: DBSCAN and CB-SMoT, is proposed. Before clustering, GPS dataset is cleaned using stop removal, missing segment interpolation and inaccurate point removal. The clustering quality measure shows that clustering dataset only after cleaning using the proposed method has more accurate clusters.

In our future work, it is expected to develop a trajectory clustering process which works automatically and gives more efficient results. For stop detection, users still need to give *minTime* value. In finding the missing points using segment interpolation, there is still needed to find a smoother interpolation method so that the estimated points are much closer to the real points. Cleaned trajectories will be used in the third stage of the research work, data enrichment, by combining the weather conditions, road speed limit and important places.

5. References

[1] Tin Lai Lai Mon, Thin Lai Lai Thein, "Design and Implementation of Smart Alert System for Reducing Road Traffic Accidents in Myanmar", *AIP Conference Proceedings 2129*, 2019.

[2] S. Kisilevich, F. Mansmann, M. Nanni, and S. Rinzivillo, "Spatio-temporal clustering", Data Mining and Knowledge Discovery Handbook, second edition, Springer, 2010.

[3] G. Andrienko, N. Andrienko, S. Rinzivillo, M. Nanni, D. Pedreschi, and F. Giannotti, "Interactive visual clustering of large collections of trajectories", *Visual Analytics Science and Technology*, pages 3 –10, 2009.

[4] L. O. Alvares, G. Oliveira, C. A. Heuser, and V. Bogorny, "A framework for trajectory data preprocessing for data mining", *Int. Conf. on Software Engineering and Knowledge Engineering*, pages 698 – 702, 2009.

ISBN 978-981-14-4787-7
Proceedings of 2020 the 10th International Workshop on Computer Science and Engineering (WCSE 2020)
Yangon (Rangoon), Myanmar (Burma), February 26- February 28, 2020, pp. 156-161

Secure Healthcare System using Blockchain Technology

HtweHtwePyone[+], KhinThan Mya

Faculty of Computer Science, University of Computer Studies (Myitkyina), Myanmar
Faculty of Computer Systems and Technologies, University of Computer Studies (Yangon), Myanmar

Abstract: Due to the popularity of crypto currencies, blockchain technology has gotten valuable for various areas. In the healthcare domain, blockchain became a key component that can drive information streams as significant and adaptable enough to enable continuous secure human services framework. So, this system is proposed as the secure healthcare system by using blockchain technology. To reduce the redundancy from each block of health data, this system proposes the modified mutual information (MMI) method as the contribution. MMI method identifies the quality of relationships between features of various natures without dispersion law limitation. And then, this system uses MD-5 (message-digest-5) hash algorithm for the hash value. Each block with hash value is transmitted through the blockchain. So, this system allows the healthcare providers to easily and securely share health data.

Key Words: Health data, Blockchain, MMI, Hash.

1. Introduction

Today, the premium and advancement of blockchain technology has been driven by the gigantic worth development of crypto-currencies and large investments of venture capital in blockchain start-ups. Crypto-currency is one of the uses of blokchain technology. There are three concepts about crypto-currency. These are blockchain, protocol and currency. In crypto-currency, the blockchain go about as a dispersed record that stores all the performed transactions. In a blockchain, new blocks are added over time. Some of blockchain's potential uses beyond crypto-currency, including for government application, healthcare, identity management and the music industry.

Blockchain can push forward the development of patient-driven medicinal services model. In this model, patients control their healthcare information. Information sharing in both patient-driven and customary models faces the absence of trust and absence of impetuses to share. For sharing data, blockchain can take care of these two issues by going about as a trust layer and presenting the motivating force instruments, for example, remunerating crypto tokens.

The proposed system aims to develop a secure healthcare system by reducing the effect of redundancy within data preprocessing stage. This system can not only reduce the irrelevant and redundant data but also decline the quantity of collinearity issues inside factor examination. To identify optimal subset of health data from the blockchain, this system proposes the modified mutual information (MMI) method. According to the MMI method, mutual information is obtained by subtracting the redundancy from the relevance. Moreover, this system uses the MD-5 hash algorithm for each block of the health data that has to be cryptographically hashed on the header of the block.

For information imparting to other human services suppliers under the user's consent, this system uploads the medical treatment data to the blockchain network. The current healthcare providers can request access to previous medical treatment and health data from the user. Both health information request and this

[+] Corresponding author. Tel.: +95 797603499
E-mail address: htwehtwepyone233@gmail.com

information access are recorded on the blockchain. Because of medical treatment history information is permanently recorded on the blockchain network, the proposed system doesn't allow the users to modify and hide those medical data. So, the integrity and trustworthiness about healthcare data are ensured in this system.

2. Related Work

In 2017, L. Xueping, Z. Juan and S. Sachin [1] presented an innovative user-centric health data sharing solution by using authorization and decentralized blockchain. This system protects privacy using channel formation scheme and enhances the identity management using the membership service that supported by the blockchain. To save the integrity of health information inside each record, a proof of integrity and validation is anchored to the blockchain network. Moreover, they embraced a tree-based information processing and batching method to handle huge data sets of personal health data that are gathered and transferred by the mobile platform.

In 2019, A. R. A. Mohammad, K. Yasar and M. C. E. Yagoub [2] presented a globally integrated healthcare record sharing architecture based on health level seven (HL7) client and blockchain. In this system, the genuine approval process is performed on a unified character the board framework, for example, the Shibboleth. Despite the fact that there are similitudes with personality the executive's frameworks, their framework includes the patient in the authorization procedure and reveals to them the characters of elements got to their health records. This system improves execution, and ensures protection and security by using blockchain and the executive's framework.

In 2019, A. A. Lukman, A. James and A. A. Emmanuel [3] proposed a method for the monitoring and securing of petroleum product distribution records in a decentralized ledger database using blockchain technology. This method is to verify the exchange of circulated records in a database and to shield records from altering, deceitful action, and debasement by the chain members. This framework demonstrated to be effective to keep up as it doesn't allow any person for records altering, however underpins understanding of 75% of members in the chain to make changes.

3. Blockchain

To set up the trust of the considerable number of components in the digital healthcare, blockchain has become as a solution. Blockchain technology uses scientific models for the circulation of encoded data through the chain of blocks. At the real time, blockchain makes the information distribution to be safe [4]. A file is represented as a block. Block can be a text file, video sample and spreadsheet. It can also be any kinds of structured data that consist of records which are storable and readable by machine. To create a chain like a process and govern the transmission of information, blocks are interconnected with nodes. Transaction is a single operation over one node. Nodes are able to communicate and transfer the data from one node to another across the network. Each node acts as a central point that is able to generate and digitally sign the transaction during the transmission process. Figure 1 shows the workflow of blockchain process.

Fig. 1: Workflow of the Blockchain Process [5]

Cryptographic hashing is also used for data transmission. By using hash algorithm, each block of data has cryptographically hashed on the header of the block. Each block contains the hash of its parent. To establish a sequence and to complete the liner list of blocks, each block in the blockchain is connected with the parent (past) block of information put away in the header: timestamp (date-time) and beginning [5].

3.1. Types of Blockchain

Blockchain includes three types that are public, federated and private blockchains. These are as follows:

- Public blockchain: Due to the permission less of public blockchain, anyone can easily validate the transaction. There is the highest level of decentralized trust because the blockchain is maintained by the public community.
- Federated blockchain: Under the leadership of a group, the federated blockchain is a permission blockchain operating. In this blockchain, the transactions may or may not be public.
- Private blockchain: Permission blockchain centralized to one governing organization is the private blockchain. In this blockchain, exchanges are approved inside and might be open lucid. This blockchain for the most part have quicker block occasions. In addition, this blockchain can process higher exchange throughput [6].

3.2. Use of Blockchain in Healthcare

In the healthcare industry, different stakeholders need to organize, access and share health records without any modification in a secure and interoperable way.

Healthcare blockchain is shown in Figure 2. Stakeholder can be practitioners, medical specialists, therapists, patients, payers, etc. To prove the authenticity of records, data provenance is essential for healthcare domain. In this situation, blockchain technology is the important for the healthcare sector. Blockchain is being implemented in different scenarios. By using blockchain, health bank provides a platform for each patient who can safely share their health data [7, 8].

4. Proposed System Design

Framework of proposed system is shown in Figure 3. Firstly, this system accepts the health data from the user. Then, this system separates the health data into each block. Then, this system eliminates the redundant data by using modified mutual information (MMI) method. By using these relevance data, this system hashed with MD5 hash function in each block of sender portion. Finally, this system produces the secure blockchain to the healthcare provider.

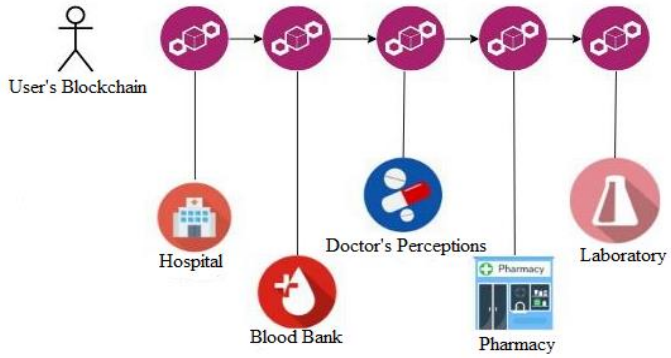


Fig. 2: Healthcare Blockchain [5]

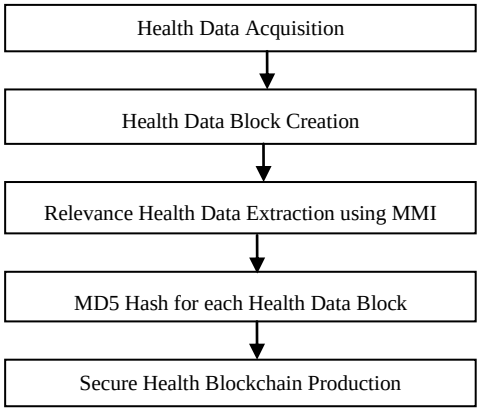


Fig. 3: Framework of the Proposed System

4.1. Mutual Information

Mutual information is a measure of statistical dependency that can determine complex relationships between features. This is a measure of the nonlinear and linear dependence between a set of features. Mutual information between two random features is a measure of the information one random variable provides about the other. If there is no dependence between the two variables, the mutual information method takes a minimum value of zero. If a strong dependence exists between the two variables, this mutual information method takes a positive method. Mutual information between two random features X and Y is as follows:

$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \quad (1)$$

where I(X;Y) is the mutual information between the two random features X and Y. The x and y represent realizations X and Y. The p(x, y) is their joint probability mass function. The p(x) and p(y) are the marginal probability mass functions.

4.2. Modified Mutual Information (MMI)

Modified mutual information (MMI) is based on identifying that the integrations of good features. Redundancy among features needs to be minimized because it is needed to maximize the joint dependency of top-ranking variables on the target features. According to mutual information, the purpose of causation-factor selection is to find a factor set S with m factors $\{x_i\}$, which have the highest mutual information value. MMI method searches the maximum relevance for satisfying factors, which approximates D(S, y) between factors x_i and class y. Maximum relevance is as follows:

$$\max D(s,y), D = \frac{1}{|s|}\sum_{x_{i\in S}} I(x_i, y) \tag{2}$$

Factors selected according to maximum relevance are likely to be highly redundant. When two factors depend highly on each other, the respective class-discriminative power would not change much if one of them were removed. To select mutually exclusive factors, MMI method adds the minimal redundancy that is as follows:

$$\min R(S), R = \frac{1}{|s|^2}\sum_{x_i, x_{i\in S}} I(x_i, x_j) \tag{3}$$

Operator Ø(D, R) combines D and R and considers the following simplest form to optimize D and R simultaneously. For features taking continuous values, which compute quantities such as the F statistic between features and the classification variable c as the score of maximum relevance that is as follows:

$$\max D(s,y), D = \frac{1}{|s|}\sum_{x_{i\in S}} F(x_i, y) \tag{4}$$

As the score of minimum redundancy, the average Pearson correlation coefficient of features is as follows:

$$\min R(S), R = \frac{1}{|s|^2}\sum_{x_i, x_{i\in S}} \left|c(x_i, x_j)\right| \tag{5}$$

MMI can also consider the distance function $d(x_i, x_j)$ for the minimum redundancy condition:

$$\min R(S), R = \frac{1}{|s|^2}\sum_{x_i, x_{i\in S}} d(x_i, x_j) \tag{6}$$

To find the causal factor relevance and redundancy, modified mutual information is taken as the basic criterion. MMI defines the relationship among different explanatory factors. MMI also finds a set of optimum casual factor that has the highest mutual information.

4.3. MD-5 Hash Function

MD-5 (Message Digest-5) processes a variable-length message into a fixed length output of 128 bits. Input message is broken up into 512 bit block. The message is padded because its length is divisible by 512. A single bit, 1, is appended to the end of the message. This is followed by many zeros to bring the length of the message up to 64 bits less than a multiple of 512. MD-5 operates on a 128 bit state, divided into four 32-bit words, denoted A, B, C and D. To modify the state, this algorithm uses each 512-bit message block. The processing of a message block consists of four stages, termed rounds. Each round is based on non-linear function F, modular addition and left rotation. These are four functions. These are as follows:

- F(B, C, D) = (B AND C) OR (NOT B AND D)
- G(B, C, D) = (B AND D) OR (C AND NOT D)
- H(B, C, D) = B XOR C XOR D
- I(B, C, D) = CXOR (B OR NOT D)

In the above four functions, a different one is used in each round.

5. Implementation of the System

The proposed secure healthcare system is implemented by using MATLAB programming language. In this system, there are two sides: the sender and receiver. This system transfers the patient healthcare data between sender and receiver. This system first loads the desired data. Then, this system splits the data into each block. From each block, this system reduces the redundant data using modified mutual information

(MMI) method. After reducing, this system encrypted this data by using MD-5 hash. Figure 4 and 5 shows the blockchain from the sender and receiver.

Fig. 4: Blockchain from the Sender

Fig. 5: Blockchain from the Receiver

6. Evaluation of the System

This system is tested different healthcare data about "Hypopharynx Carcinoma" and "Schwannoma" disease. To evaluate the performance of modified mutual information (MMI), this system uses the Univariate and Multivariate analysis methods. Mutual information rate is between "0" and "1". Table 1 shows each features about "Hypopharynx Carcinoma" and "Schwannoma" disease.

Table 1: Features about "Hypopharynx Carcinoma" and "Schwannoma" Disease

Features	Code	Diagnosis
History of present illness	HPI	Schwannoma
Lt sided weakness than Rt side	LTSide	Schwannoma
Lt eye blurred vision	LTEye	Schwannoma
Rt eye normal	RTEye	Schwannoma
Can't walk well	Cwalk	Schwannoma
Odynophegia 1.5 months	ODY	Hypopharynx Carcinoma
Cough +	COU	Hypopharynx Carcinoma
change of voice +	CV	Hypopharynx Carcinoma
Congestion of Pulmonary Disease	CPD	Hypopharynx Carcinoma
Healed, dry 4 cm diameter wound	HDW	Schwannoma

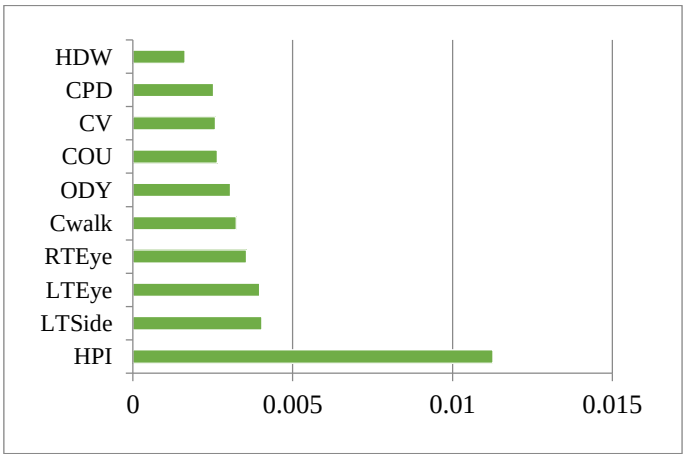


Fig. 6: Univariate Analysis based on MMI

Univariate analysis based on modified mutual information (MMI) is shown in Figure 6. Multivariate analysis with MMI for "Schwannoma" diagnosis and "Hypopharynx Carcinoma" diagnosis are shown in Figure 7 and 8.

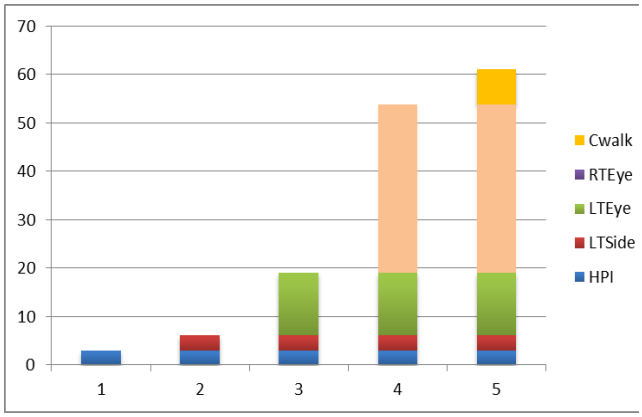


Fig. 7: Multivariate analysis with MMI for "Schwannoma" diagnosis

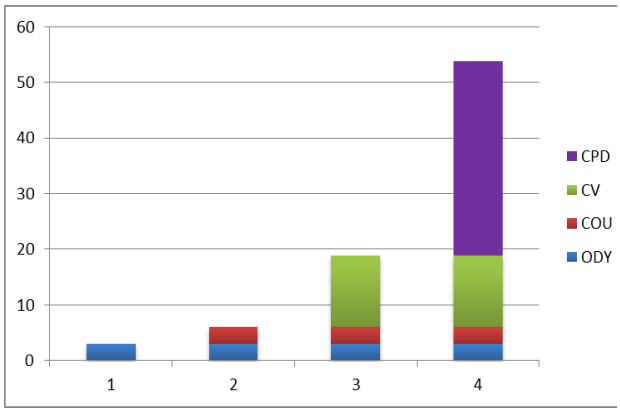


Fig. 8: Multivariate analysis with MMI for "Hypopharynx" diagnosis

7. Conclusion

In the patient-centric and traditional models, health data sharing faces the lack of trust and incentives to share. By using blockchain technology, the proposed system solved these problems. This system acts as a trust layer for sharing healthcare data. Moreover, blockchain that is produced from the system can be the bridge for the integration of medical device data and healthcare internet of things; the healthcare and lifestyle data collected by wearable devices can be critical for correct diagnosis since there is a lack of a proper way for a physician to access the patient-generated data.

8. References

[1] L. Xueping, Z. Juan and S. Sachin, "Integrating Blockchain for Data Sharing and Collaboration in Mobile Healthcare Application, IEEE, 2017.

[2] A. R. A. Mohammad, K. Yasar and M. C. E. Yagoub, "Fusing Identity Management, HL7 and Blockchain into a Global Healthcare Record Sharing Architecture", *International Journal of Advanced Computer Science and Applications*, vol. 10,no. 6, 2019, pp. 630-636.

[3] A. A. Lukman, A. James and A. A. Emmanuel, "Crypto Hash Algorithm-Based Blockchain Technology for Managing Decentralized Ledger Database in Oil and Gas Industry", *Multidisciplinary Scientific Journal*, vol. 2, 2019, pp. 300-325.

[4] G. Sylvester, *Blockchain*, Food and Agriculture Organization of the United Nations and the International Telecommunication Union, Bangkok, 2019.

[5] D. Rakic, "Blockchain Technology in Healthcare", *In Proceedings of the 4th International Conference on Information and Communication Technologies for Ageing Well and e-Health*, 2018, pp. 13-20.

[6] G. J. Katuwal, S. Pandey and M. Hennessey, " Applications of Blockchain in Healthcare: Current Landspace& Challenges", *arXiv*, 2018.

[7] S. Yaqoob, M. M. Khan and R. Talib, "Use of Blockchain in Healthcare: A systematic Literature Review", *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 5, pp. 644-653, 2019.

[8] S. Sourabh, "Healthcare Blockchain Leads to Transform Healthcare Industry", *International Journal of Advance* Research, *Ideas and Innovations in Technology*, IJARIIT, 2018.

ISBN 978-981-14-4787-7
Proceedings of 2020 the 10th International Workshop on Computer Science and Engineering (WCSE 2020)
Yangon (Rangoon), Myanmar (Burma), February 26- February 28, 2020, pp. 162-167

The Analysis of Landslide Based on Geographic Information System in Mon State, Myanmar

Chaw Chaw Khaing [1], Thin Lai Lai Thein [2 +]

[1, 2] University of Computer Studies, Yangon

Abstract. Land use change can increase or decrease landslide susceptibility in the mountainous areas. In the hilly and mountainous part of Mon State, Myanmar, land use change has taken place due to land extractions and rock extractions. These activities can worsen the slope susceptible to sliding due to mostly the wounding of the mountain. So, every year take place the landslide in monsoon season. The objective of this study is to define the landslide risk areas in support of development planning, monitoring, and control of unstable areas. In this study, the mapping of landslides using Sentinel 2, research on their combination for discerning historical landslides in the raining season. Landslide samples were obtained from the old landslide, road structure from the topo map and slope can get form digital elevation model (DEM)). Layers were analyzed and the average weighted score was applied to calculate every 9 classes to predict the landslide. Overlay, geoprocessing and geostatistical techniques in geographic information systems (GIS) were used to discriminate these weighted subclasses into landslide features 6 levels of risk areas. Landslides in the Paung township and Melamine at Mon State which showed the prone area with the study.

Keywords: Landslide, Geographic Information System, Sentinel 2

1. Introduction

Landslides are one of the disasters and that occurred pagoda it was on the top of the mountain and near the highway road. A lot of landslides occurred in near river regions and in the mountain regions. Many major landslides always occurred in mining region, caused by human activity. This study has six major factors causing landslides have been analyzed in GIS, the weighted overlay method with each six factors combination to get susceptibility landslide map. Sentinel 2 image can evaluate the normalized difference vegetation index (NDVI) and the normalized difference water index (NDWI) value. In this study, histogram values are classified by 9 with Natural Breaks (Jenks) method. And land cover land use map used unsupervised classification with K-Means Algorithm.

2. Study Areas

The Paung hazard prone area is located in the south region of Myanmar. The Study area covers an area of about 603,332,510.665499 Square Meters between longitudes 97 32´E and 97 33´E and between latitudes16 27´N and 16°32´N. Fig 3. The entire area is located near the Thanlyin river and Gulf of Martaban. There are a lot of streams and the range of mountain regions. Land use land cover changes by the local company are extracts the land. This problem caused slope deformation and landslide. Fig. 1.

3. The Data Source

The data used in the study mainly include a topographic map for road pattern, a geological map for lithology, landslide reports for old landslide, sentinel 2 satellite 10 m resolution image. Sentinel-2 s an Earth observation mission from the Copernicus programmed that acquires optical imagery with high spatial

[+] Corresponding author. Tel.: + 95 01 610655; fax: +013-610-633
E-mail address: chawchawkhaing@ucsy.edu.mm

resolution (10 m) and systematic image, over land and coastal waters. This constellation mission has two twin satellites, Sentinel-2A and Sentinel-2B. It supports a large range of services and applications such as agricultural monitoring, emergency management, land cover classification or water quality. the European Space Agency (ESA) developed Sentinel-2 and operated, and Airbus Defense and Space (Airbus DS) manufactured satellites by a consortium led.

Fig. 1: Lower Myanmar, Paung and Melamine Township, Mon State

Table 1: Sentinel 2 satellite sensor specifications

Band Name	Band Width (mm)	Resolution	Purpose
Band 2	65	10	Blue
Band 3	35	10	Green
Band 4	30	10	Red
Band 5	15	20	Vegetation Classification
Band 6	15	20	Vegetation Classification
Band 7	20	20	Vegetation Classification
Band 8	115	10	Near infrared

The metrological data (numeric data) can't get the update from the Department of Meteorology and Hydrology (Myanmar) and also USGS, now can get up to July, 2019. The study of the landslide was on 9 August 2019.

3.1. Featured Based on Digital Elevation Model

3.1.1. Slope

Slope degree is one of the most frequently-used factors in assessing landslide susceptibility. It has a great influence on slope stability and is directly related to the different types of mountain hazards. A slope is the rise or fall of the land surface. A slope is easy to recognize in a hilly area. Slope represents the rate of change of elevation for each digital elevation model (DEM) pixel. It represents the steepness of the surface and is symbolized into three classes that are shown using color saturation (brightness). Figure 2.

3.1.2. Aspect

The slope aspect is defined as the direction of the terrain surface, such as north, northeast, west, southwest. It identifies the downslope direction of the maximum rate of change in value from each pixel to its neighbors. It can be thought of as the slope direction. The values of the output raster will be the compass direction of the aspect, represented by a hue (color). Figure 3.

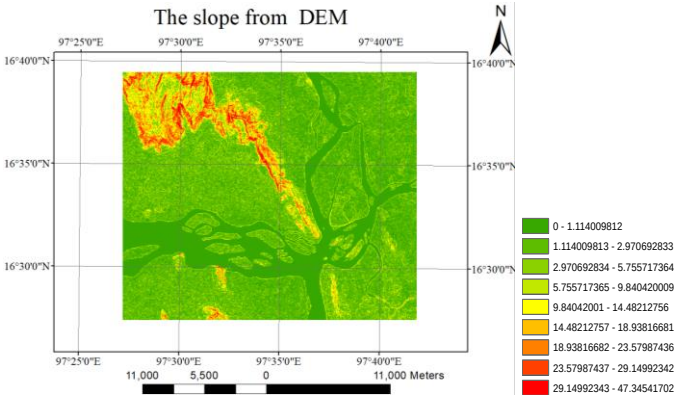


Fig. 2: The Slope

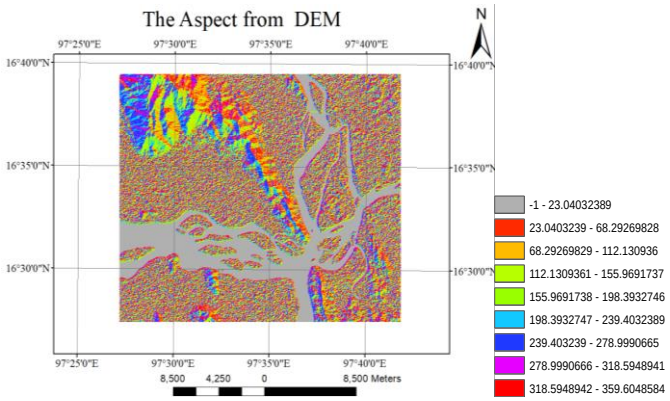


Fig. 3: The Aspect

3.1.3. Contours

Contours are lines that connect locations of equal value in a raster dataset that represents continuous phenomena such as elevation, temperature, precipitation, pollution, or atmospheric pressure. The line features connect cells of a constant value in the input. Contour lines are often generally referred to as isolines but can also have specific terms depending on what is being measured. Some examples are isobars for pressure, isotherms for temperature, and isohyets for precipitation. Fig 4.

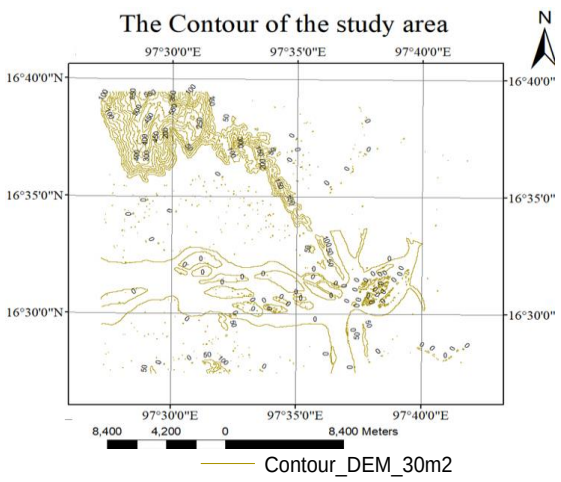


Fig. 4: Contour of the study Area

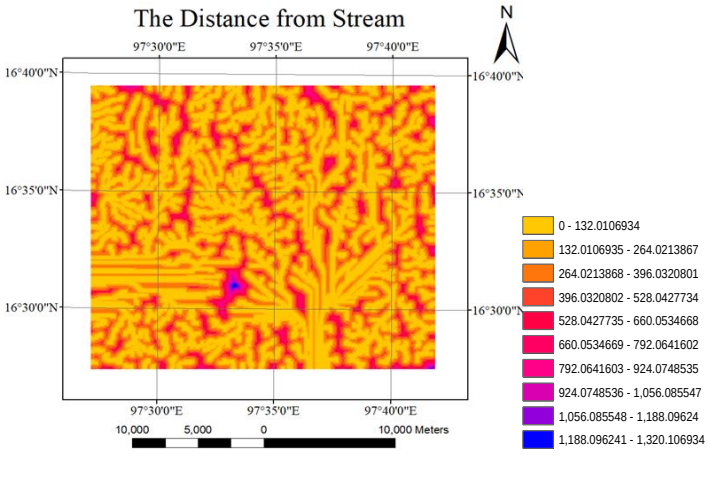


Fig. 5: Distance from Stream of the study Area

3.1.4. Distance of Stream

Stream Distance calculates from DEM using flow direction, flow accumulation, stream order. And to get the nearest distance stream use the Euclidean distance algorithm. The shortest distance to a source is determined, and if it is less than the specified maximum distance, the value is assigned to the cell location on the output raster. Fig 5.

3.2. Featured Extract from Topo Map

Road Distance calculates the road shape from topo map. And to get the nearest distance road also use the Euclidean distance algorithm. Fig 6.

3.3. Feature Extracted from Remote Sensing Image

Remote sensing images, geometric correction images are used to extract land cover and utilization information through object-based classification methods. This process used unsupervised classification and K-means algorithm. Then different types of land covers are classified, tree, forest, agriculture, waterbody, road, open land and wetland.

3.3.1. Distance of Stream Normalized Difference Vegetation Index (NDVI)

The normalized difference vegetation index (NDVI) is a simple graphical indicator that can be used to analyse remote sensing images, images accessed by NDVI can interpret chlorophyll colour, so vegetation area is whether or not in the accessed images. Fig 7.

$$NDVI = \frac{(NIR - Red)}{(NIR + Red)}$$

Where NDVI stands for normalized difference vegetation index, NIR stand for nearest infrared band and Red stand for Red Band. NIR band and red band describe in table1, sentinel 2 satellite sensor specifications.

$$NDWI = \frac{(Green - NIR)}{(Green + NIR)}$$

Where NDWI stands for normalized difference water index, Green stands for green band and NIR stands for near infrared band. The green band describes in table 1, sentinel 2 satellite sensor specifications. Fig8. The classification results obtain vegetation index, NDVI index, NDWI index, road index, and stream index.

3.4. Feature Based on Geological Map

3.4.1. Lithology

The lithology and solid sources are upper Paleozoic (mainly C.P) Moulmein limestone [1]. The Nitsoil (NT) soil type is the World Reference Base for Soil Resources (WRB), this soil is a deep, red, well- drained soil. It contains more than 30% of a clay and structure is a blocky. This soil type correlate with the Kandic Alfisols, Ultisols and Inceptisols of the United States Department of Agriculture soil taxonomy.

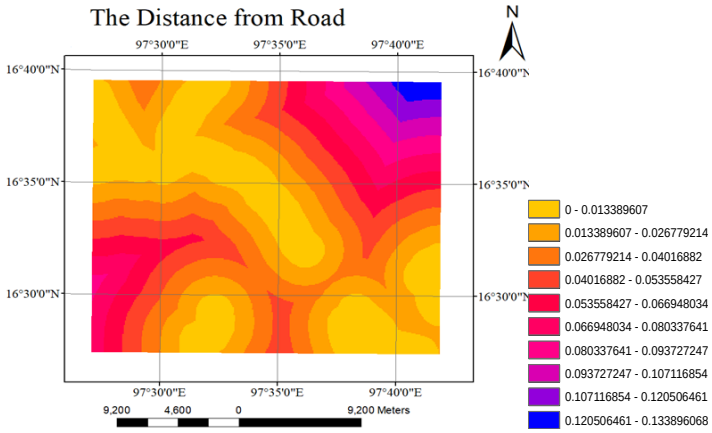


Fig. 6: The road distance map

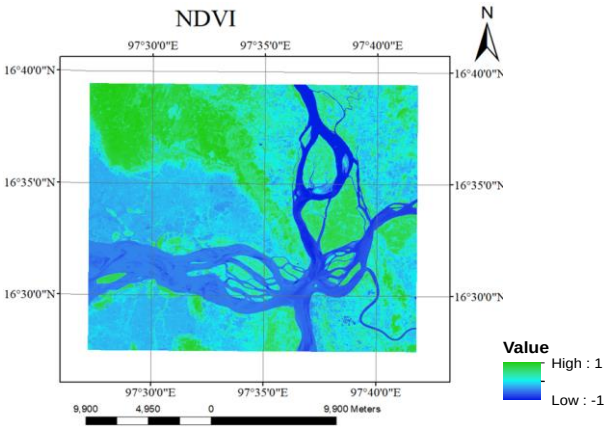


Fig. 7: NDVI

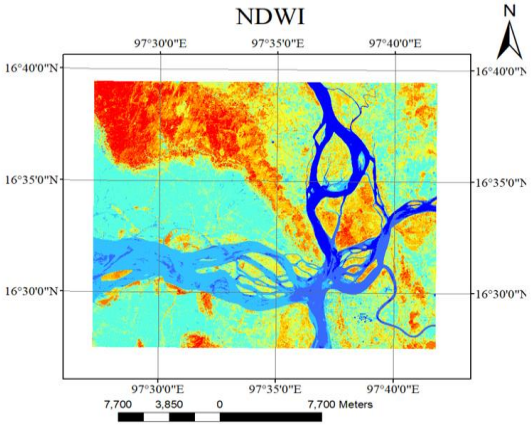


Fig. 8: NDWI

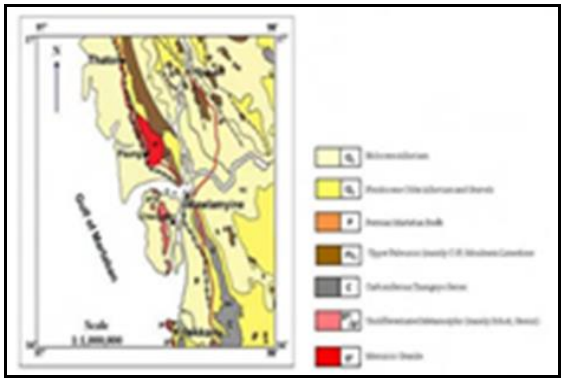


Fig. 9: Geological Map

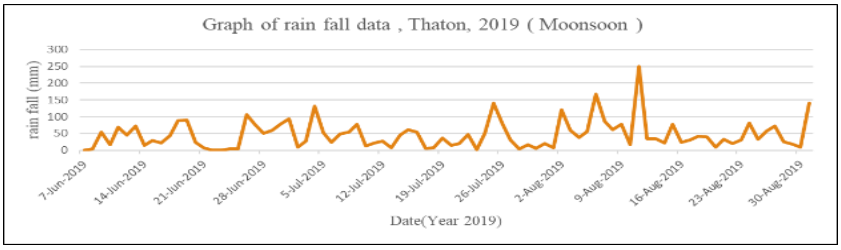


Fig. 10: Rainfall data

3.5. Feature Based on Metrological Data

Rainfall is an important factor in triggering landslides because it reduces soil suction and increases the pore-water pressure in soils [2]. This graph showed the landslide in the month June, July and August, during the monsoon season. This update data got from the local department, department of metrological and hydrology, Yangon, Myanmar. This graph showed the maximum rainfall day, this day occurred the landslide event.

4. Experiment Result

Deep learning algorithm, long short-term memory (LSTM) have four states, input gates, forget gates, cell units, output gates and learning the length of the input sequence data controls the number of the historical data points in the recursive connection. By the grid search method, the input sequence length is set to 6 with parameter is input sequence length and loss function [3].

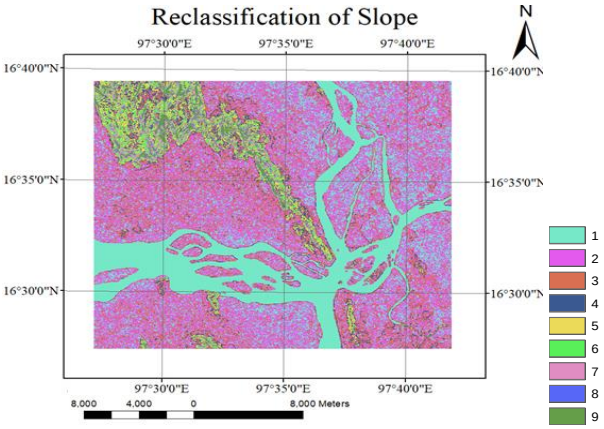


Fig. 11: Reclassification of Slope

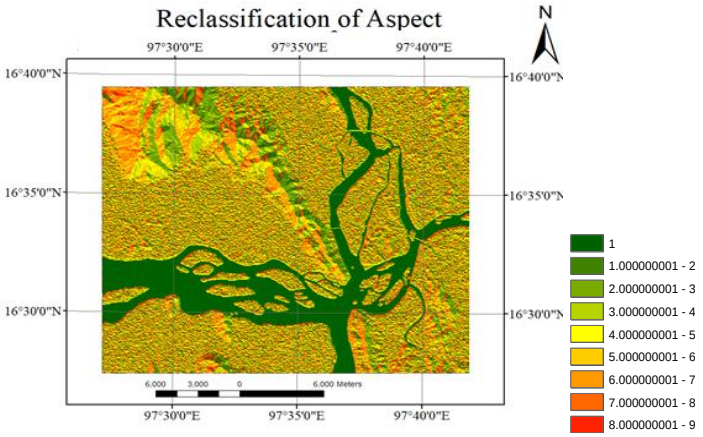


Fig. 12: Reclassification of Aspect

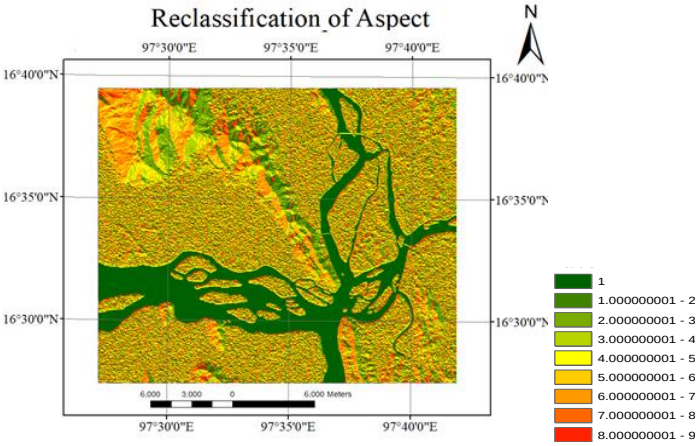


Fig. 13: Reclassification of Aspect

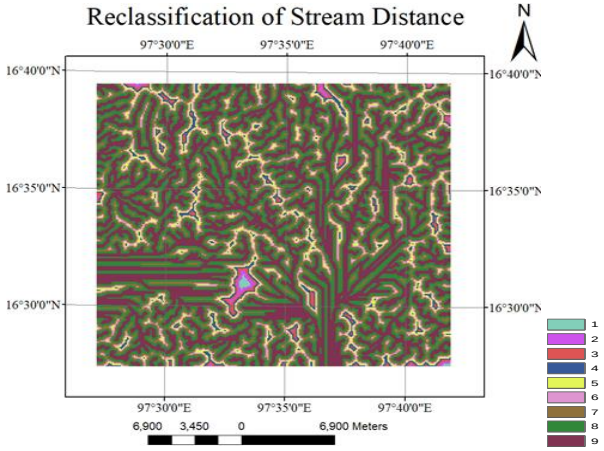


Fig. 14: Reclassification of Stream distance

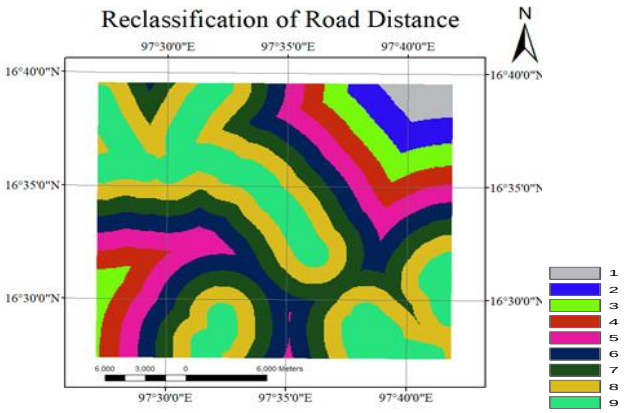


Fig. 15: Classification of Road distance

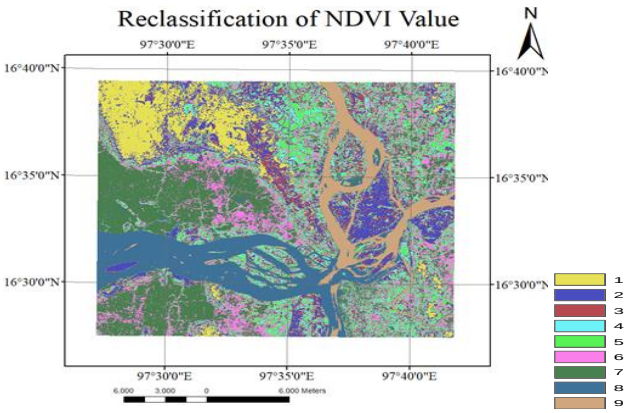


Fig. 16: Classification of NDVI value

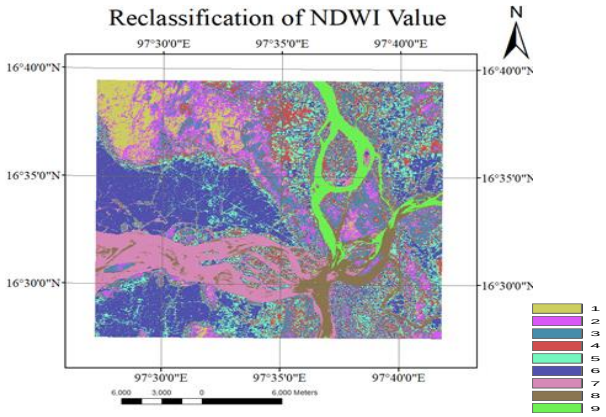


Fig. 17: Reclassification of NDWI value

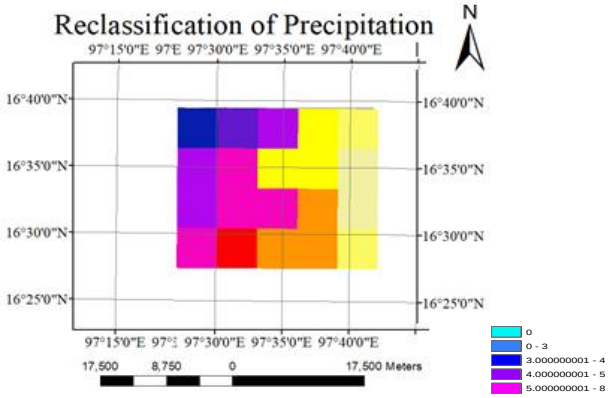


Fig. 18: Classification of precipitation

5. Result Output

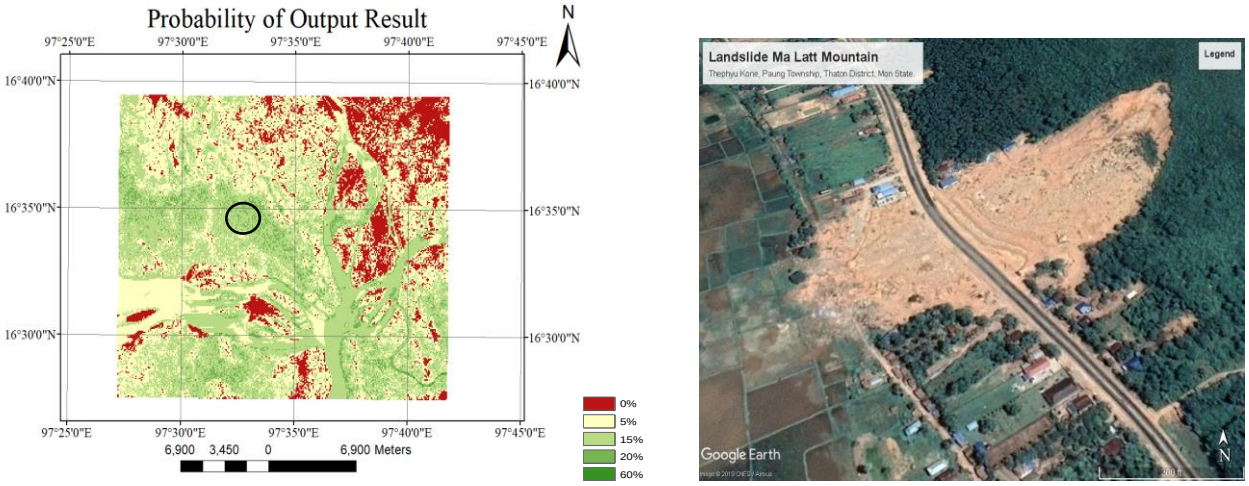


Fig. 19: Landslide Potential Map

6. Conclusion

The Paung township is a land extraction area of the Mon state. On the mountain and foot, the mountain has a lot of extraction. Due to this extraction, the natural environment decay and along with the hill wetland area increase on the mountain. Therefore, at least two days rainfall, on the top of wetlands are flowed under the mountain. This includes rock and clay, they are flowed. And then the foot of the mountain is very near the high way road, the stream, and river. This image can interpret the landslide prone area. The rainfall data from USGS, TRMM can get rainfall surface up to June, but this event took previous month of August. This historical data is proposed to solve the problem of landslide susceptibility. Geological data, geographic information, high-resolution remote sensing images, hydrological data can use features of LSTM model and predict the next disaster event during the moon soon, heavy rainfall two days. This historical product the model for real time monitoring system for landslide prevention. The other researchers construct the classification models with BPNN, SVM, and DT, which are also applied for comparisons with the LSTM model in landslide susceptibility assessments. The results of their study showed that the SVM model (72.87%) had better accuracy than the BPNN (62.03%) and DT model (60.42%). The LSTM model (81.18%) outperformed SVM in prediction accuracy [3].

7. Acknowledgements

I would like to thank my Supervisor, Dr. Thin Lai Lai Thein, Professor, the University of Computer Studies, and I would like to express thanks to Dr. Myint Myint Sein, GIS Lab, the University of Computer Studies, Yangon, for allowing me to develop of study and giving me general guidance during the period of my study. I would also like to express my respectful gratitude to Dr. Sabai Phyu, Professor, Dean of the Ph.D. 11th batch, the University of Computer Studies, Yangon, and Saya U Win Ko for their patient supervision, tenderness, encouragement and providing me. I would like to express my respectful gratitude to all my teachers for their encouragement. I also thank my friends from the Ph.D.11th batch for their co-operation and encouragement.

8. References

[1] Min Min Khaing. Petrological Analysis of the Granitoid Rocks from Pangon Area, Paung Township, Mon State. Dagon University Commemoration of 25th Anniversary Silver Jubilee Research Journal 2019, Vol.9, No.2 384.

[2] Tuhua Ma, Changjiang Li, Zhiming Lu, BaoxinWang. An effective antecedent precipitation model derived from the power-law relationship between landslide occurrence and rainfall level. *Geomorphology 216 (2014) 187–192.*

[3] Liming Xiao, Yonghong Zhang and Gongzhuang Peng. "*Landslide Susceptibility Assessment Using Integrated Deep Learning Algorithm along the China-Nepal Highway", Sensors, MDPI.*

ISBN 978-981-14-4787-7
Proceedings of 2020 the 10th International Workshop on Computer Science and Engineering (WCSE 2020)
Yangon (Rangoon), Myanmar (Burma), February 26- February 28, 2020, pp. 168-172

Service Management Strategy for CDN

Xinhua E[1+], Binjie Zhu[2], Hui Zhang[2] and Yanjun Shi[2]

[1]Faculty of Information Technology, Beijing University of Technology, Beijing, 100124, China
[2] China Mobile Group Beijing Company Limited, Beijing, 100007, China

Abstract. A content distribution network is a kind of overlay network to improving Internet access speed for clients. The management of cache server of content distribution network is an important issue. This paper studies the edge server resource management strategy of streaming media service in the multi-service CDN. A management strategy was proposed in this paper. It can manage the edge server resources in the effective under the premise of guaranteeing QOS, and improve resource utilization in the CDN.

Keywords: CDN, Overlay network, service management

1. Introduction

Content distribution network is a kind of overlay network, which was used to improve the user access latency [1]. The traditional Internet access model is the C / S access patterns, in where the user directly access to a server directly from the servers to obtain the requested contents. C / S model has many shortcomings. When the load of the server was high, the server will be congestion. The user was far from the server, resulting in greater user response delay.

A layer of edge servers were added in content distribution network between the users and the servers [2]. Contents were distributed through a certain algorithm to be sent to the appropriate edge server. When client requesting a content, the request were redirected to the nearest edge server. If the edge server has the content requesting, then the edge servers service the client. If it has not the content, the request was redirected to other servers until it finds the content of user requests.

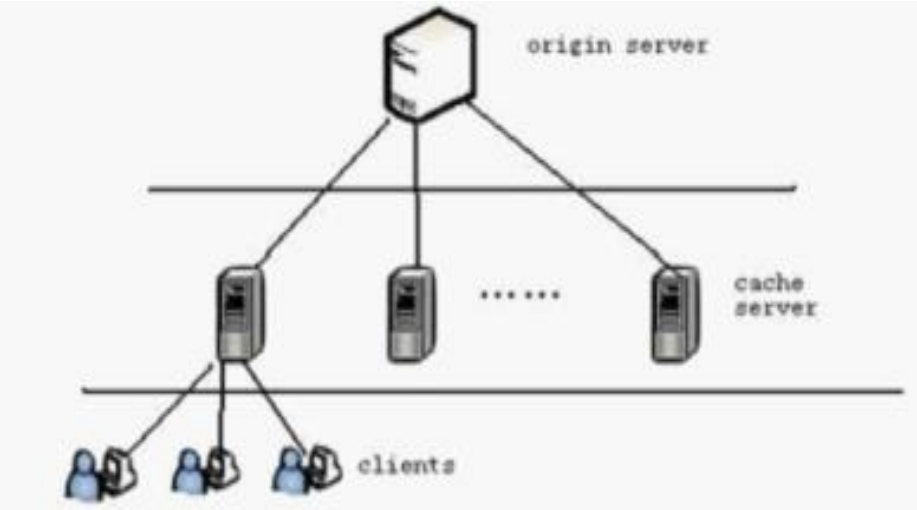


Fig. 1: CDN

+ Corresponding author. Tel.: + 86 13811324402
E-mail address: Elson1900@163.com

2. Background and Motivation

The CDN architecture is one of the key issues for content delivery network. Different network architecture needs to be designed based on different delivery demands. The network architecture can directly affect content delivery, request routing and other strategies.

P2P-based content delivery network can use three fusion ways, namely, direct overlay fusion architecture, core architecture and complete peer-to-peer architecture.

The design concept of overlay fusion architecture is to overlay the loose coupling of the CDN and P2P system. Generally, the overlay fusion architecture is divided into two layers, namely, CDN and P2P, of which, the CDN system architecture is divided into two layers. Therefore, this is a three-layer architecture including the CDN management layer, the CDN edge server layer and the user P2P layer. The CDN management layer manages the entire system. The CDN edge server layer is at the edge of the network, and the CDN edge server pushes contents to the network edge near the users by caching content objects. Several user nodes close to an edge server form a peer-to-peer network, generally an unstructured network. The user nodes not only can obtain content objects from peer nodes, but also can access contents objects from CDN edge server nodes.

The nodes in a CDN in a complete peer-to-peer architecture can be turned into peer nodes by two ways. CDN is generally divided into management layer and core functional layer. The management layer is responsible for optimizing and managing the function of the entire CDN system. The core functional layer is responsible for realizing specific content storage, service, search and other core functions. The first method is to turn all nodes in the management layer into peer nodes in the P2P network, that is, both of nodes in the management layer and nodes in the functional layer are peer nodes in the P2P network. The second method is to distribute the functions of the CDN management layer into nodes at each core layer. In this way, each node has a part of management functions and the management function can be realized through the cooperation of the management node modules in each of the peer nodes. The first method is complete peer-to-peer at the network level and the second one is complete peer-to-peer at the function level.

The core transformation architecture constitutes the P2P layer of the core functional layer through the P2P transformation of the core functional layer of CDN. In this architecture, the management mode of the management layer is not changed. Instead, only peer-to-peer of the core functional layer is implemented. The goal of peer-to-peer is to improve the overall performance by enhancing collaboration at the core functional level. This approach does not require the participation of the clients. The clients only need to make a content request and access content services, so that the overall network resources can be centrally managed and billed.

P2P-CDN is a special CDN, introduced the idea of P2P into CDN. CDN's caches were organized by P2P. It can share resources between caches. Figure 2 is a CDN, the principle: the user redirects the user's request through the system-oriented nearest cache server. As shown in figure 2.The user's request is redirected to the nearest edge server redirected system. If the edge server has the request contents, it will response the request. If the edge server has not the request content, the edge server finds the request resources and pulls to the local server. The edge servers form a resource sharing network [3].

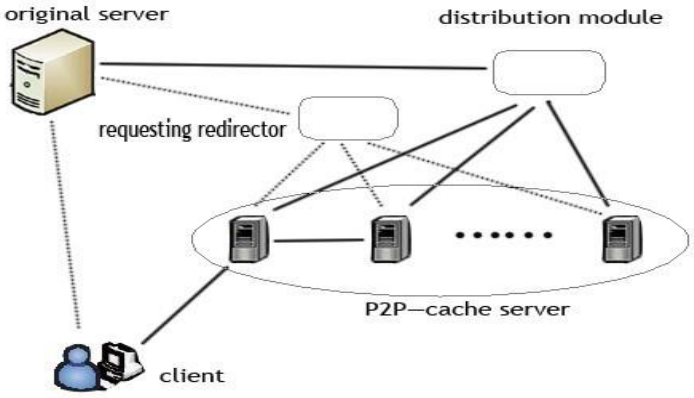


Fig. 2: Multi-service P2P-CDN

In order to ensure the QOS of cache server, most of streaming media CDN will limit the maximum number, and then request as a queue according to the rules FCFS (First Come First Servers). When a user requests occupy server resources, the server will always be for this user until users quit the requests. Management strategy for the request take a video file as a full-service, there are two problems. First, the resources were limited, while the maximum number of users in services is limited .Second, the speed that the content download from the server faster than the speed of the user view, so when the user has the exit behavior, the download the file at this time far beyond the length of time users need to watch. So it is waste of resources of the server.

This paper designs a server resource management strategies, the performance be enhanced in two ways: First, in the case of the same server resources, enhance the number of simultaneous users of the service. Second, the services were effective management avoiding the waste of server resources.

3. CDN Resource Management Strategy

The access frequency between each section within a program is a difference. Client typically demand weather continue a movie after looking at the beginning of time, and according to the contents of it form an overall evaluation of the film, and then decide whether to continue . The client's behavior characteristics analysis in paper [4], described by the probability of a user exit during the viewing behavior.

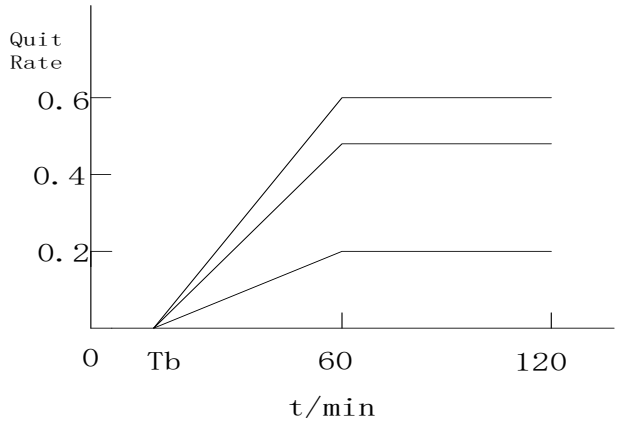


Fig. 3: Quit rate of a content

The download speed was taken as the service speed, the watching speed as the speed of service consumption. Service speed is great than service consumption rate. If take a video file download as a service, we can design a strategy for the server to efficiently manage bandwidth resources.

In order to ensure the QOS of cache server, most of streaming media CDN will limit the maximum number, and then request as a queue according to the rules FCFS (First Come First Server). When a user requests occupy server resources, the server will always be for this user until users quit the requests.

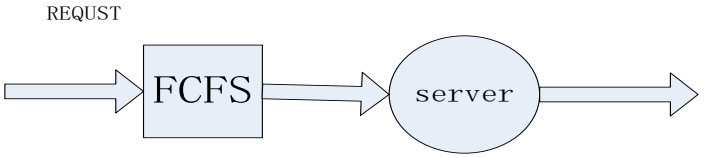


Fig. 4: First come first service

The basic idea of this patent is to create a new rule to manage the request queue on the server resources. Suppose there are R requests coming, the rule is that each request queue was divided into N sub-requests. The first M sub-requests were selected from the R queuing requests as a group. Take the first sub-requests from the first request in the queue group, and then the first sub-requests from the second request in the queue group. And then traverse from 1 to M, taking the second ub-request. Until the last traverse 1 to M, remove the last N sub-requests. Handle the request in accordance with the queuing algorithm to form a new queue, and the server's queue according to the new service.

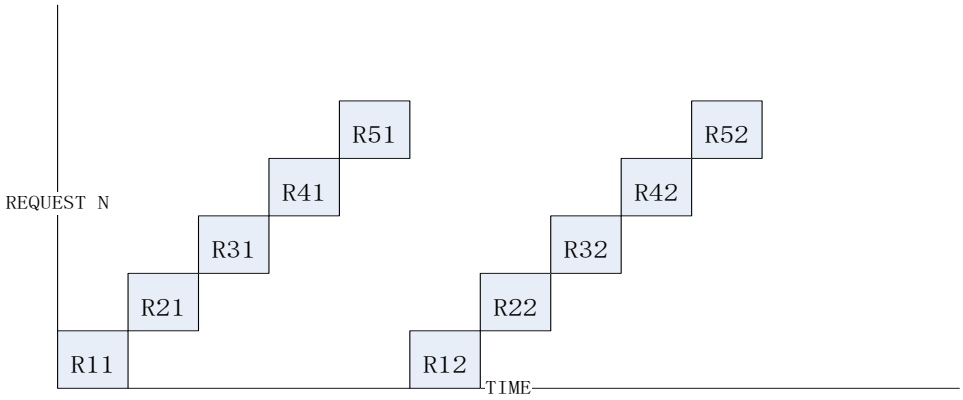


Fig. 5: First come portion first service

4. Analysis

For contrast analysis, a scenario was supposed. The large of all the video files were 120 Minutes. The storage spaces for every minute file were 1MB.The download speed is 0.1MB/S. The download time for one video is 12 Minutes.

The service time of FCFS strategy is 12 Minutes. The average service time of FCPFS strategy is 6.6 Minutes.

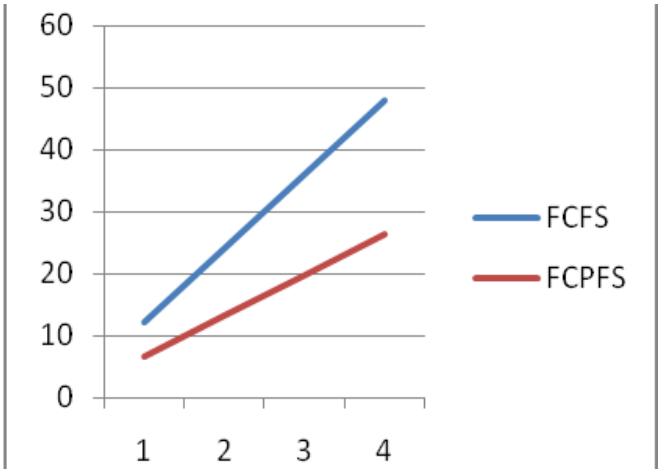


Fig. 6: Service time

Each edge server was taken as M|M|n|0 queue system with arrival rate 0.2.The number of the server in queue system is 2.FCFS strategy as situation A. FCPFS as situation. In the situation A，

$$\rho_0 = \frac{1}{1+\rho+\frac{\rho^2}{2!}} = 0.158 \tag{1}$$

$$\rho_1 = \rho\rho_0 = 2.410 * 0.158 = 0.381 \tag{2}$$

$$\rho_2 = \frac{\rho^2}{2!}\rho_0 = 0.459 \tag{3}$$

The contrast of Pr as figure 7.The contrast of Q as figure 8.

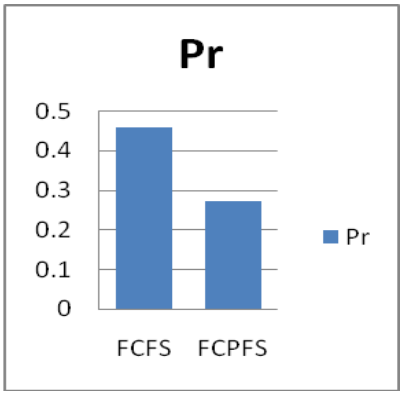


Fig. 7: Pr of strategy

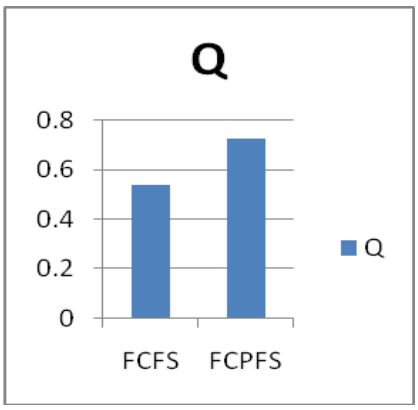


Fig. 8: Q of strategy

5. Conclusion

The management of cache server of content distribution network is an important issue. This paper studies the edge server resource management strategy of streaming media service in the multi-service CDN. A management strategy was proposed in this paper. It can manage the ES resources in the effective under the premise of guaranteeing QOS, and improve resource utilization in the CDN. Form the contract, we can conclusion that the proposed strategy was effective.

6. References

[1] Sanaa Sharafeddine, Karim Jahed, Omar Farhat, et al. Failure recovery in wireless content distribution networks with device-to-device cooperation[J]. Computer Networks, 2017, 128(9): 108-122.

[2] Rami Halloush, Hang Liu, Lijun Dong, et al.Hop-by-hop Content Distribution with Network Coding in Multihop Wireless Networks[J]. Digital Communications and Networks, 2017, 3(1): 47-54.

[3] Xiaoying Zheng, Ye Xia.Optimizing network objectives in collaborative content distribution [J]. Computer Networks, 2015, 91(14): 244-261.

[4] Uttam Mandal, Pulak Chowdhury, et al.Energy-efficient networking for content distribution over telecom network infrastructure[J]. Optical Switching and Networking, 2013, 10(4): 393-405.

ISBN 978-981-14-4787-7
Proceedings of 2020 the 10th International Workshop on Computer Science and Engineering (WCSE 2020)
Yangon (Rangoon), Myanmar (Burma), February 26- February 28, 2020, pp. 173-180

A Content Sharing System Using Dynamic Fog Consisting of Peer-to-Peer Terminals and Its Simple Evaluation

[1]Takuya Itokazu and [2]Shinji Sugawara[+]

[1,2]Chiba Institute of Technology, Japan

Abstract. This paper proposes an efficient content sharing system by forming a *Dynamic Fog* with multiple Edge terminals that make up a Peer-to-Peer network and logically constructing a Cloud-Fog-Edge hierarchical structure. In general, by placing a Fog server between Cloud and an Edge terminal, redundant traffic between Cloud and Edge nodes can be reduced, and the delay time required for providing contents or transmitting some kind of directions from Cloud can be shortened by shortcuts using Fog. In this research, it is novel that the role of a Fog server is cooperatively played by multiple Edge terminals, instead of a relatively high-performance dedicated Fog server. The proposed system is evaluated by simple computer simulations in order to show its potential effectiveness.

Keywords: Cloud-Fog-Edge hierarchy, content sharing, Dynamic Fog, Peer-to-Peer.

1. Introduction

With the spread of high-speed and high-quality communication networks in recent years like 5G and the performance improvement of terminal devices such as PCs, tablet computers, and smartphones, the environment where a large number of users access a wide variety of contents such as images, videos, sounds, haptics, texts, and so on, and the environment where they are searching their favorite contents over the network or providing them each other have already been established. In addition, the effective use of abundant network resources including contents has been achieved.

In previous research, we proposed a content sharing system based on a hybrid of Peer-to-Peer network and Multi-Cloud, which assumed the contribution of user terminals, and clarified its effectiveness [1], [2], [3]. A system that executes a service combining a Cloud and Edge terminals generally has a two-layer structure of a Cloud layer and an Edge layer. When Edge terminals constitute a Peer-to-Peer network as shown in Fig. 1, or when each Edge terminal is working as a high-performance client with partial function of a server, the Edge terminals are connected to the system via the access network.

On the other hand, in recent years, in consideration of the situation around the Edge terminals, their regional characteristics, and user group preferences, etc. in order to provide services that respond to the needs that have been finely differentiated, in order to improve their response speed, or in order to offload network traffic and computational load, etc., computer systems with a three-layer structure in which an intermediate layer subsystem called Fog is arranged between the Cloud layer and the Edge layer, have been introduced and applied in various fields [4]. Such a three-layer structure is also sufficiently effective for content sharing, because it is expected that the index information which is unique to each region will be collectively shared in the Fog layer, and this causes that the contents will be exchanged quickly. However, if the Fog function is arranged as a dedicated server, the maintenance cost of the server will be caused in addition to the Cloud's. On

[+] Corresponding author. Tel.: +81 47 478 0393.
E-mail address: shinji.sugawara@it-chiba.ac.jp

top of that, it is necessary to keep this server even when there is no use opportunity, and it is difficult to make effective use of computational resources.

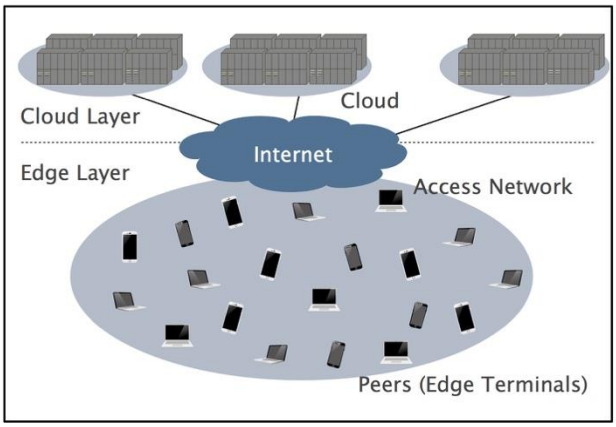


Fig. 1: Cloud-Edge Terminal: Two-Layer Structure.

We have already proposed to construct a content sharing system that combines the Peer-to-Peer-based system with a Cloud [1]. So, in this paper, the fundamental challenge is to dynamically compose a Fog function of intermediate layer by peers including mobile Edge terminals.

If it is possible to prepare a wide variety of mobile terminals, which compose a Peer-to-Peer network and share contents by coordinating with preliminarily prepared processes running on the Cloud, and if they in rotation and autonomously provide a necessary scale of virtual computing function of a Fog server according to each terminal's performance and situation, a very flexible and efficient content sharing system can be realized.

2. Related Works

Although there are not many similar proposals about the composition of a Fog function with dynamic nodes, the following two pieces of research have similar elements with ours.

In [5], a large number of terminals in the network are clustered into several groups, and each cluster is considered as one virtual machine. By allocating an appropriate load to each cluster, the authors aim to achieve a large-scale computation similar to the cloud.

In this research, they focus on dynamically switching and combining terminals in cluster generation, integrating them with a graph-theoretic approach, and mainly on appropriately distributing the computational load.

The purpose of our research is to share contents on the premise that the Edge terminals move, join, and leave the network autonomously, so there is a clear difference between this research and ours. On top of this, the system configuration is only terminals in this research, and there is no physical Cloud-Edge hierarchy like ours.

In [6], the system has a clear physical Cloud-Fog-Edge hierarchy, and the main idea is that Edge terminals can dynamically generate clusters according to the situation and be used for offloading of the Fog server.

It can be said that each Edge terminal is a part of the Fog server, but the Fog serer physically exists, and this research is very different from ours in the point that the Edge terminals are not combined according to the situation. The purpose of this research also seems to be efficient clustering with appropriate combinations of Edge terminals.

Unfortunately, the above two pieces of research have not been fully evaluated their effectiveness, and the objectives and main techniques are different from ours.

3. Proposed Method

As mentioned above, we have been conducting research on content sharing using Peer-to-Peer networks so far, and have proposed a method that aims to reduce storage capacity and network load, and improve content acquisition rate by adding a Cloud system.

And in recent years, in addition to simple systems consisting of Cloud and a group of peers which are a large number of user terminals, Fog computing and Edge computing aiming to shorten response time to users, to reduce network load, or to offload the server's processing running in the Cloud have attracted attention.

However, if Fog is used as an always-on server, it will be necessary to consider the costs of installation, management, and maintenance of the server, in addition to the Cloud's cost. In this case, it seems to be more difficult for us to evaluate the effectiveness of the system, such as what kind of cost we should assume as a system operating side, or who actually spends the cost.

Therefore, only two components of the system are used as before: The Cloud and a number of user terminals (Edge terminals) that make up the Peer-to-Peer network. In addition to performing Edge processing, peers dynamically configure a virtual Fog function in this paper, so a three-tier structure of Cloud-Fog-Edge can be logically established.

Although the resources equivalent to the server prepared by the system operating side are allocated in the required number of places near the terminals in normal Fog computing, in this research, this server function is virtually realized mainly by multiple user terminals that can be operated as high-performance Edge terminals, and they play a role of the Fog.

In such a system configuration, as shown in Fig. 2, the system provider can operate the system using only the resources allocated on the Cloud and the applications stored in each user Edge terminal. This makes the system configuration simple and can reduce the operating cost.

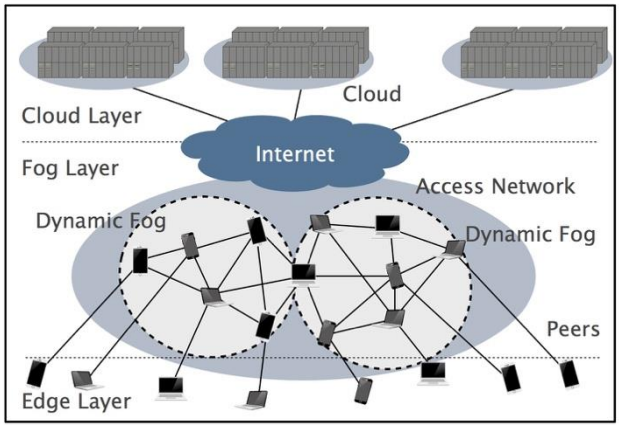


Fig. 2: System Configuration of Proposed System.

Although it is logically a Cloud-Fog-Edge three-layer system, it is physically a two-layer structure as shown in Fig. 1, and because a dynamic Fog can change its scale as necessary, there is no need to dedicate Fog resources during its unused periods of time and the system configuration can be efficient for both the system provider and the users.

Because dynamic Fog is logically configured, plural number of Fog functions are constructed at the same time for different purposes of content sharing such as with different shared genres, in different languages, or about regional differences, and each Edge terminal can join in the different plural number of Fog functions at the same time, too.

The core of the proposed system is the dynamic Fog part, and even if multiple Edge terminals move and repeatedly join and leave the Fog, the leader terminal, to be described below, cooperated with Cloud, tries to avoid the loss of the index information, which is needed for contents searching and allocation, and makes the system stable for the content sharing. The outline of the operation is described below. Note that we assume each Edge terminal on the access network, that can be a component of a Dynamic Fog in the initial state, is registered in the Cloud preliminarily.

1. The terminal to be the leader is determined by the Cloud according to the specification of each terminal.

2. A dynamic Fog of appropriate scale is formed by the notification from the leader terminal to other terminals.
3. Periodic report including connection confirmation from each terminal to the leader using the access network is started.

From then, a content held by each terminal and the index information related to the content are shared and uploaded among terminals configuring Fog, when it is needed.

If there is a concern that the contents or index information may be lost due to the movement of the terminal, terminal's leaving the network, or the failure of the terminal, etc., the index information is appropriately transferred to another terminal, or copied and evacuated to the Cloud at the discretion of the leader.

General terminals connect to the terminals that configuring Fog, and always exchange its location information and possessing-contents information, and sometimes send content requests. In response to a content search request regardless of its sender, index information can be obtained by pure Peer-to-Peer system like behavior of the terminals configuring the Fog, and finally the requested content is sent from its owner to the requesting peer (terminal). If the requested content and its index information cannot be found in the Fog, the content is inquired to Cloud, and an attempt is made to obtain it from another Fog nearby or a Fog existing in another access network.

All the terminals move from one location to another, and join or leave the network at any time, and the terminals that make up the Fog jointly grasp them, and share index information and optimally relocate the contents.

4. Evaluation

In this paper, we evaluate the working effectiveness of the content sharing system described above, using a computer simulation under the simple scenarios and assumptions. In order to verify the effect of Dynamic Fog, we compare the results both when using Dynamic Fog and when not using it, under the environment where the Fog server is used. The simulator is made based on The One [7]. The main conditions set in the simulation are as follows.

- The nodes that make up the simulation are three types: Edge nodes, a Fog server, and Cloud.
- Up to 2,000, 3,000 and 4,000 Edge nodes (the average number of the nodes during the simulation is 1,250, 1,550, and 1,850, respectively) are randomly placed on 2 km × 2 km plane in the initial state.
- Each node (considered Edge or user) is roughly equally divided into three categories based on its attributes (generation, etc.), and each node passes through 14 points which are randomly selected from 250 points on the plane, consisting of 125 points unique to each category and 125 points common to all categories.
- After that, each node moves to a specific point (assuming the starting or end point of high-speed transportation, such as a train station or an airport, etc.), and then disappears. Also, every 10 minutes, Edge nodes equivalent to 1% of total number of the nodes on the plane are newly added from this point.
- Each Edge node can exchange contents and queries by direct communication between nodes, and communication via base station of access network.
- Every 2 minutes, 20% of all Edge nodes are randomly selected, and each of these Edges makes a request for a single content to the Fog server via the base station of the access network. Then the location information of the requested content can be obtained from the Fog index information.
- There are 100 or 140 titles of content in total and these are all stored in the Cloud at any time, however the Fog and Edge nodes do not have any content in the initial state.
- Content capacity has a variety of 250 kB, 500 kB, 1 MB, 3 MB, 5 MB, and 10 MB, and their percentage of the total number of content titles is 25%, 25%, 15%, 15%, 10%, and 10%, respectively.
- The capacity that can hold the contents of each node is 100 MB, however it is 300 MB only when the node serves as a part of Dynamic Fog. If it is necessary to hold a copy of a content exceeding this capacity, the one that had possessed at the earliest is deleted by FIFO.
- If an Edge node sends a content request to the Fog, and the Fog does not hold neither the content nor its index information, the Fog server gets the content from the Cloud, caches it in its own storage, and sends

it to the requested Edge node. This Edge node is memorized by the Fog server as an Edge node having the content.

- The storage capacity of the Fog server is 600 MB and if it is necessary to hold a copy of a content exceeding this capacity, delete some with LFU.
- The content requested by an Edge node is determined by the following procedure.
  1. With a probability of 50%, it is selected from all the titles of the contents.
  2. Otherwise (for the case of the remaining 50%), it is selected according to the popularity.
  3. Each titles of the contents is classified into one of the three categories of top, middle, and low popularity, and each percentage of the total number of all titles is 20%, 30%, and 50%, respectively.
  4. If the content title to be requested is selected according to the popularity, it is selected randomly from the categories of top, middle, and low popularity with the probability of 50%, 30%, and 20%, respectively.
- Some Edge nodes are converted to Dynamic Fog nodes so that the Edge nodes that make up Dynamic Fog occupy 10% of the entire Edge nodes once every 20 minutes. However, those Edge nodes should be decided taking care not to be close to each other. If there are no suitable Edge nodes, less than 10% will be acceptable.
- This simulation stated above is continued for 12 hours, and the number of communications (inquiries to Cloud, inquiries to the Fog server, communication through base stations between Edge nodes, and direct communication between Edge nodes), and the ratio of them are calculated. The calculated values are the average ones of simulation runs of 50 times.

Table 1: The main parameters of the simulation

Parameters	Values
Maximum number of Edge nodes	2,000, 3,000, 4,000
Average number of Edge nodes	1,250, 1,550, 1,850
Number of content titles	100, 140
Storage capacity of Edge node (normal state)	100 [MB]
Storage capacity of Edge node (when serving as a part of Dynamic Fog)	300 [MB]
Simulation area	2 [km] ×2 [km]
Capacity of a content title	250 [kB], 500 [kB], 1 {MB}, 3 [MB], 10 [MB]
Storage capacity of Fog server	600 [MB]
Simulation period	12 [hours]
Number of simulation-runs	50

The main parameters included in the explanation of the simulation stated above is illustrated in Table 1.

The result of the simulation is shown in Figs 3 through 7.

Fig. 3 and 4 show the ratio among communications of an Edge node to Cloud via a base station, to Fog server via a base station, to another Edge node via a base station, and to another Edge node directly, for content sharing by the procedure explained above, without using Dynamic Fog.

The difference between Fig. 3 and 4 is only the total number of sharing contents titles. The former shows in the case of 100 titles, and the latter shows in that of 140 titles. As the number of sharing contents titles increases, the ratio of content sharing between Edge nodes decreases, and that between an Edge node and Cloud or the Fog server increases. This is a reasonable result, because each Edge node has a small storage capacity and there is a limit to the number of copies of contents titles that can be stored, so if the total number of contents titles increases, content sharing should rely on Cloud or the Fog server.

Turning to Fig. 5 and 6 based on the result above, when Dynamic Fog is used together with the Fog server, it can be seen that the communication between Edge nodes increases rapidly whether via a base station or not.

This indicates that Dynamic Fog has been able to almost replace the content sharing that relied on the Fog server, and it is considered that Dynamic Fog works effectively under the conditions given this time.

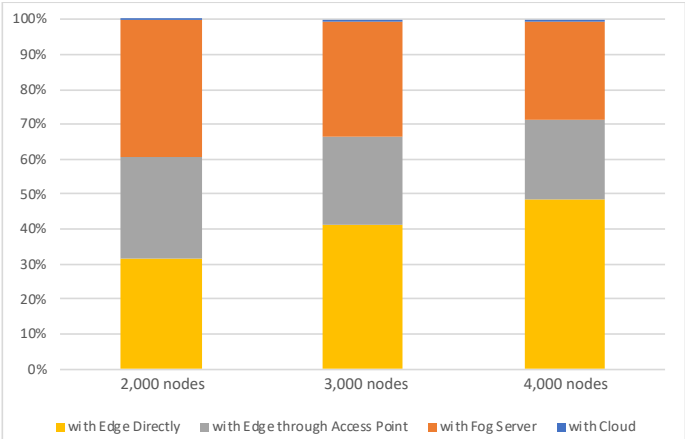


Fig. 3: Ratio of Communications (without Dynamic Fog, the number of contents titles is 100).

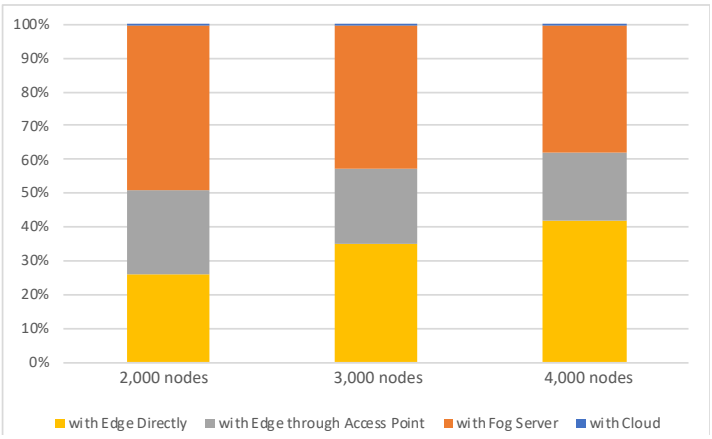


Fig. 4: Ratio of Communications (without Dynamic Fog, the number of contents titles is 140).

The difference between results of Fig. 5 and 6 is the total number of contents titles to be shared, so if this increases from 100 to 140, the usage ratio of Cloud and the Fog server increases just the same as the results when only the Fog server is used, shown in Figs. 3 and 4.

From Figs. 3 through 6, it can be seen that as the number of Edge nodes increases, the ratios of communications between Edge nodes without using a base station tend to increase. This is because, from the view point of each Edge node, the probability that there exist many other Edge nodes nearby that can cooperatively share contents is getting higher, and the content exchange which does not have to rely on the distant Edge nodes, the Fog server, nor Cloud to be connected via a base station is increasing.

Figure 7 shows the total number of communications in the case of using Dynamic Fog. As the number of Edge nodes increases, the number of communications increases, which is also a reasonable result.

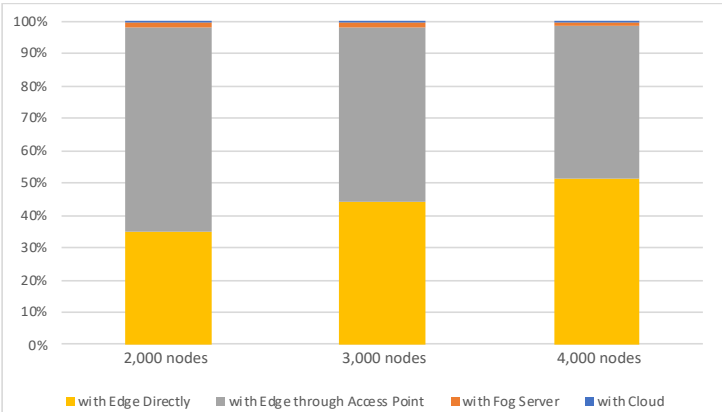


Fig. 5: Ratio of Communications (with Dynamic Fog, the number of contents titles is 100).

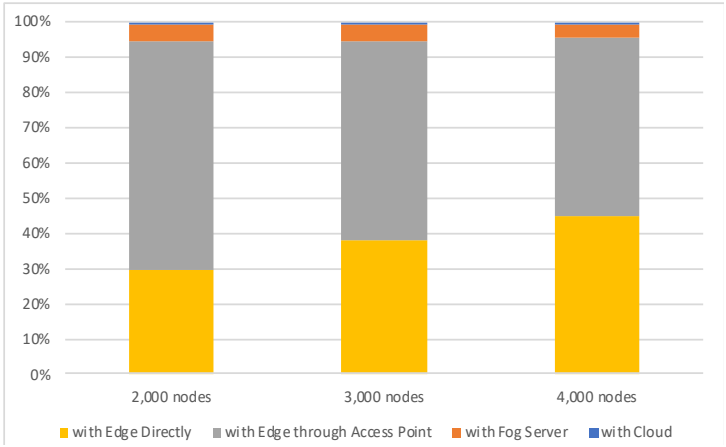


Fig. 6: Ratio of Communications (with Dynamic Fog, the number of contents titles is 140).

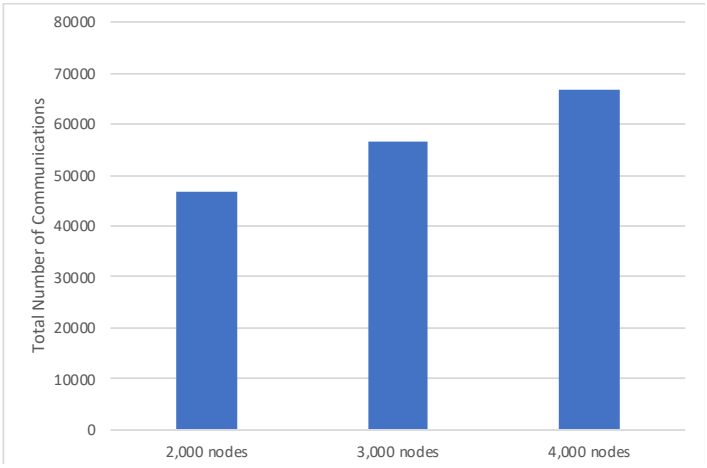


Fig. 7: Total Number of Communications.

From the above, it is clarified that the use of Dynamic Fog, at least under the condition set this time, could potentially replace the conventional Fog server function.

5. Conclusion

In this paper, we proposed an efficient content sharing system, which makes Dynamic Fog by the plural number of Edge nodes forming a Peer-to-Peer network. In Dynamic Fog, normal Edge nodes cooperate each other and replace the function of the conventional Fog server. Therefore, the content sharing system as a whole does not need to prepare dedicated resources for the Fog function, and flexible operation can be expected.

In order to confirm the effectiveness of the proposed method, we investigated by computer simulation how Dynamic Fog can replace the function of Fog server under a simple scenario. As a result, comparing the environment where only the Fog server was used and that where both were co-existed, it was revealed that the communication for content sharing largely depending on the Fog server in the former environment, could be replaced by Dynamic Fog in the latter one. From the above, it is considered that the proposed method has a potential effectiveness at least in the environment set this time.

As a future work, it is necessary to define the behavior of Dynamic Fog in more detail and to build a system that can be expected to operate more efficiently. It is also an important issue to evaluate the proposed method more accurately under the condition that is more realistic than that we set this time.

6. References

[1] M. Tomimori, S. Sugawara, “Content Sharing Method Using Expected Acquisition Rate in Hybrid Peer-to-Peer Networks with Cloud Storages,” Int. J. Space-Based and Situated Computing, Vol. 7, No. 4, pp. 187-196, Feb. 2018.

[2] T. Murakami, S. Sugawara, “An Upload and Download Time Shortening Method for Multi-Cloud Content Sharing

Systems," in Proc. 2019 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW 2019), May 2019.

[3] T. Murakami, S. Sugawara, "Multi-cloud System for Content Sharing Using RAID-Like Fragmentation," Proc. The 22nd International Conference on Network-Based Information Systems (NBiS-2019), Advances in Intelligent Systems and Computing, 1036, pp. 49-60, Springer, Sep. 2019.

[4] S. Sugawara, "Survey of Cloud-based Content Sharing Research: Taxonomy of System Models and Case Examples," IEICE Trans. Commun., vol. E100-B, no. 04, pp. 484-499, April 2017.

[5] N. Song et al., "Fog Computing Dynamic Load Balancing Mechanism Based on Graph Repartitioning," China Communications, Vol. 13, Issue 7, pp. 156-164, Mar. 2016.

[6] Y. Li et al., "Dynamic Mobile Cloudlet Clustering for Fog Computing," Proc. International Conference on Electronics, Information and Communication (ICEIC 2018), Jan. 2018.

[7] The One, the opportunistic network environment Simulator: https://akeranen.github.io/the-one/

Author Index

A
Adnan Ahmed Khan 122
B
Binjie Zhu 168
C
Chang-Sung Sung 46
Chaw Chaw Khaing 162
D
Dim Lam Cing 63
E
Eaint Mon Win 136
F
Fang Changjie 84
Feipei Lai 46
H
Hiroaki Nishino 68
Hiroyuki Kawai 103
Hitoshi Watanabe 114
Htet Htet Naing 52
HtweHtwePyone 156
Hui Zhang 168
I
Iku Kitanosono 68
Imran Rashid 122
J
JongWon Kim 73
Junghan Ha 130
K
Khaing Yee Mone 57
Khin Mar Soe 36, 63
Khin Myo Myat 16
Khin Thandar Nwet 1
Khin Than Mya 108, 156
Khine Khine Oo 27
Kiran Khurshid 122
Koki Makino 90
Koudai Houga 103
L
Liu Ziming 84
Lwin May Thant 22
M
May Aye Khine 136
May Zin Oo 108
Muwook Pyeon 130
Muhammad Ahmad Rathore 73
Muhammad Haroon Siddiqui 122
Myint Myint Moe 27
Myint Myint Sein 11
P
Pau Suan Mung 141
Pingguo Huang 114
R
Raungrong Suleesathira 95
S
Sabai Phyu 22, 141
Sandar Win 79
Seona Park 130
SeungHyung Lee 73
Shinji Sugawara 173
Si Si Mar Win 16
Ssu-Ming Wang 46
T
Takanori Miyoshi 103
Takuya Itokazu 173
Thanh-Tho Quan 6
Thazin Myint Oo 36
Thein Yu 1
Thepchai Supnithi 36
Thin Lai Lai Thein 79, 162
Thi-Kim-Anh Vo 6
Tin Lai Lai Mon 151
Tien-Dung Phan 6
Toshiro Nunome 90
Toshiyuki Haramaki 68
W
Wai Mar Hlaing 11
Wonwoo Jung 130
X
Xinhua E 168
Y
Yang Chen 46
Yanjun Shi 168
Yasunori Kawa 103
Ye Kyaw Thu 36
Yoshifumi Hisanaga 146
Yoshio Yamamoto 57
Yutaka Ishibashi 108, 114
Yuichiro Tateiwa 146
Z
Zin May Aye 52