Advances in Intelligent Systems and Computing 388

Thi Thi Zin
Jerry Chun-Wei Lin
Jeng-Shyang Pan
Pyke Tin
Mitsuhiro Yokota *Editors*

Genetic and Evolutionary Computing

Proceedings of the Ninth International Conference on Genetic and Evolutionary Computing, August 26–28, 2015, Yangon, Myanmar – Volume II

Springer

Advances in Intelligent Systems and Computing

Volume 388

Series editor

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland
e-mail: kacprzyk@ibspan.waw.pl

About this Series

The series "Advances in Intelligent Systems and Computing" contains publications on theory, applications, and design methods of Intelligent Systems and Intelligent Computing. Virtually all disciplines such as engineering, natural sciences, computer and information science, ICT, economics, business, e-commerce, environment, healthcare, life science are covered. The list of topics spans all the areas of modern intelligent systems and computing.

The publications within "Advances in Intelligent Systems and Computing" are primarily textbooks and proceedings of important conferences, symposia and congresses. They cover significant recent developments in the field, both of a foundational and applicable character. An important characteristic feature of the series is the short publication time and world-wide distribution. This permits a rapid and broad dissemination of research results.

More information about this series at http://www.springer.com/series/11156

Thi Thi Zin · Jerry Chun-Wei Lin
Jeng-Shyang Pan · Pyke Tin
Mitsuhiro Yokota
Editors

Genetic and Evolutionary Computing

Proceedings of the Ninth International Conference on Genetic and Evolutionary Computing, August 26–28, 2015, Yangon, Myanmar – Volume II

Springer

Editors
Thi Thi Zin
Faculty of Engineering
University of Miyazaki
Miyazaki
Japan

Pyke Tin
Faculty of Engineering
University of Miyazaki
Miyazaki
Japan

Jerry Chun-Wei Lin
School of Computer Science and Technology
Harbin Institute of Technology Shenzhen Graduate School
Shenzhen
China

Mitsuhiro Yokota
Faculty of Engineering
University of Miyazaki
Miyazaki
Japan

Jeng-Shyang Pan
College of Information Science and Engineering
Fujian University of Technology
Fuzhou
China

ISSN 2194-5357 ISSN 2194-5365 (electronic)
Advances in Intelligent Systems and Computing
ISBN 978-3-319-23206-5 ISBN 978-3-319-23207-2 (eBook)
DOI 10.1007/978-3-319-23207-2

Library of Congress Control Number: 2015946740

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

Preface

This volume composes the proceedings of the Ninth International Conference on Genetic and Evolutionary Computing (ICGEC 2015), which was hosted by University of Computer Studies, Yangon and was held in Yangon, Myanmar on 26–28, August, 2015. ICGEC 2015 was technically co-sponsored by Springer, Ministry of Science and Technology, Myanmar, University of Computer Studies, Yangon, University of Miyazaki in Japan, Kaohsiung University of Applied Science in Taiwan, Fujian University of Technology in China and VSB-Technical University of Ostrava. It aimed to bring together researchers, engineers, and policymakers to discuss the related techniques, to exchange research ideas, and to make friends.

93 excellent papers were accepted for the final proceeding. Three plenary talks were kindly offered by: Professor Chin-Chen Chang (IEEE Fellow, IET Fellow, Feng Chia University, Taiwan), Professor Yutaka Ishibashi (IEICE Fellow, Nagoya Institute of Technology, Japan), and Professor Jun Murai (Keio University, Japan). Prof. Jun Murai is known as the "Internet samurai" and, in Japan has also been called "the father of the Internet in Japan".

We would like to thank the authors for their tremendous contributions. We would also express our sincere appreciation to the reviewers, Program Committee members and the Local Committee members for making this conference successful. Finally, we would like to express special thanks for the financial support from University of Miyazaki, Japan in making ICGEC 2015 possible.

June 2015

Thi Thi Zin
Jerry Chun-Wei Lin
Jeng-Shyang Pan
Pyke Tin
Mitsuhiro Yokota

Conference Organization

Advisory Committee Chairs

Hiromitsu Hama	(Honorable Professor) Osaka City University, Japan
Pyke Tin	(Former Rector) University of Computer Studies, Myanmar
Bin-Yih Liao	Kaohsiung University of Applied Sciences, Taiwan
Fujio Imai	University of Miyazaki, Japan
Jun Murai	Keio University, Japan

Advisory Committee Members

Thomas Coughlin	USA
R.M. Phatarfod	Australia
Thein Oo	Myanmar ICT Development Cooperation, Myanmar
Takashi Toriu	Osaka City University, Japan
Toshiaki Itami	University of Miyazaki, Japan
Masahiro Hamano	Asia SEED, Japan
Akinori Nishihara	Tokyo Institute of Technology, Japan

General Chairs

Kyaw Zwa Soe	Ministry of Science and Technology, Myanmar
Tatsuo Suganuma	University of Miyazaki, Japan
Vaclav Snasel	VSB-Technical University of Ostrava, Czech Republic

Program and Steering Committee Chairs

Jeng-Shyang Pan	Fujian University of Technology, China
Xin Huan Jiang	Fujian University of Technology, China
Thi Thi Zin	University of Miyazaki, Japan

Local Organization Committee Chairs

Mie Mie Mie Thet Thwin	University of Computer Studies, Yangon, Myanmar
Aye Myint	Yangon Technological University, Myanmar

Invited Session Chairs

Junzo Watada	Waseda University, Japan
Ching-Yu Yang	National Penghu University, Taiwan
Shu-Chun Chu	Flinders University, Australia

Publication Chairs

Mitsuhiro Yokota	University of Miyazaki, Japan
Jerry Chun-Wei Lin	Harbin Institute of Technology Shenzhen Graduate School, China

Finance Chairs

Mie Mie Khin	University of Computer Studies, Mandalay, Myanmar
Soe Soe Khaing	University of Technology (Yatanarpon Cyber City), Myanmar

Social Event Chairs

Myint Myint Than	Myanmar Computer Federation, Myanmar
Nay Chi Lai Lai Thein	ICTTI, Myanmar
Nang Saing Mon Khan	University of Computer Studies, Yangon, Myanmar

Program Committee Members

Noriyuki Hayashi	University of Miyazaki, Japan
Koichi Tanno	University of Miyazaki, Japan
Saw Sandar Aye	University of Information Technology, Myanmar
Moe Pwint	University of Computer Studies, Mandalay, Myanmar
Win Aye	Computer University, Mandalay, Myanmar
Thin Thu Naing	Computer University, Taunggyi, Myanmar
Aung Win	University of Technology (YCC), Myanmar
Khin Mar Lar Htun	Computer University, Hinthada, Myanmar
Myat Thidar Mon	University of Information Technology, Myanmar
Thandar Thein	Computer University, Aubin, Myanmar
Ei Ei Hlaing	Computer University, Taungoo, Myanmar
Nilar Thein	Computer University, Meikhtila, Myanmar
Myint Myint Khaing	Computer University, Pinlon, Myanmar
Nan Saw Kalayar	Computer University, Taunggyi, Myanmar
Than Nwe Aung	University of Computer Studies, Mandalay, Myanmar
Khun Oo	Myanmar Computer Federation, Myanmar
Tianhua Liu	Shenyang Normal University, China
Chang-Tsun Li	University of Warwick, UK
Sheau-Dong Lang	Central Florida University, USA
Feng-Cheng Chang	Tamkang University, Taiwan
Yuh-Yih Lu	Minghsin University of Science and Technology, Taiwan
Yueh-Hong Chen	Far East University, Taiwan
Fa-xin Yu	Zhejiang University, China
Chin-Chen Chang	Feng Chia University, Taiwan
Alexander Gelbukh	National Polytechnic Institute, Mexico
A. Fuster-Sabater	Institute of Applied Physics, Spain
Aurora Trinidad Ramirez Pozo	Federal University of Parana, Brazil
Brijesh Verma	Central Queensland University, Australia
Chien-Chang Hsu	Fu-Jen Catholic University, Taiwan
Chung-Huang Yang	National Kaohsiung Normal University, Taiwan
Chang-Shing Le	National University of Tainan, Taiwan
Chao-Chun Chen	Southern Taiwan University, Taiwan
Cheng-Hsiung Hsieh	Chaoyang University of Technology, Taiwan
Chian C. Ho	National Yunlin University of Science Technology, Taiwan
Djamila Ouelhadj	University of Portsmouth, UK
Enrique Herrera-Viedma	University of Granada, Spain
Gheorghita Ghinea	Brunel University, UK

Georgios Ch. Sirakoulis	Democritus University of Thrace, Greece
Yue Li	Nankai University, China
Huey-Ming Lee	Chinese Culture University, Taiwan
Isabel Jesus	Institute of Engineering of Porto, Portugal
Isabel L. Nunes	Universidade Nova Lisboa, Portugal
Jorge Nunez Mc Leod	Universidad Nacional de Cuyo, Argentina
Jiun-Huei Ho	Cheng Shiu University, Taiwan
Jose Alfredo F. Costa	Federal University (UFRN), Brazil
Kazumi Nakamatsu	University of Hyogo, Japan
Kwok-Yan Lam	Tsinghua University, China
Lily Lin	China University of Technology, Taiwan
Lotfi Ben Romdhane	Lotfi Ben Romdhane, Tunisia
Massimo De Gregorio	Istituto di Cibernetica, Italia
Mario Koeppen	Kyushu Institute of Technology, Japan
Michal Wozniak	Wroclaw University of Technology, Portugal
Mohsen Askari	University of Technology Sydney, Australia
Mauricio Papa	University of Tulsa, USA
Ramin Halavati	Sharif University of Technology, Iran
Shu-Chuan Chu	Cheng-Shiu University, Taiwan
Sebastian Ventura	University of Cordoba, Spain
Selva S. Rivera	Universidad Nacional de Cuyo, Argentina
Yung-Jong Shiah	Kaohsiung Medical University, Taiwan
Michael N. Vrahatis	University of Patras, Greece
Xiao-Jun Zeng	University of Manchester, UK
Xiao-Zhi Gao	Helsinki University of Technology, Finland
Yongjian Hu	South China University of Technology, China
Yu-lung Lo	Chaoyang University of Technology, Taiwan
Yusuke Nojima	Osaka Prefecture University, Japan
Zhigang Zeng	Huazhong University of Science and Technology, China
Xinpeng Zhang	Shanghai University, China
Jun-Bao Li	Harbin Institute of Technology, China
Yongping Zhang	Hisilicon Technologies Co., China
Albert B. Jeng	Jinwen University of Science and Technology, Taiwan
Wang Lidong	National Computer Network Emergency Response technical Coordination Center of China, China
Tsung-Che Chiang	National Taiwan Normal University, Taiwan
Show-Jane Yen	Mining Chuan University, Taiwan
Jung-San Lee	Feng Chia University, Taiwan
Wen-Yang Lin	National University of Kaohsiung, Taiwan
Julio Cesar Nievola	Pontificia Universidade Catolica do Parana, Brazil
Stefanos Gritzalis	University of the Aegean, Greece
Jerry Chun-Wei Lin	Harbin Institute of Technology Shenzhen Graduate School, China

Martine De Cock	Ghent University, Belgium
Ruqiang Yan	Southeast University, China
Yung-Fa Huang	Chaoyang University of Technology, Taiwan
Yau-Hwang Kuo	National Cheng kung University, Taiwan
Chia-Chen Lin	Providence University, Taiwan
Li Yao	University of Manchester, UK
Jae C. Oh	Syracuse University, USA
Estevam R. Hruschka Jr.	Federal University of Sao Carlos, Brazil
Heitor Silverio Lopes	Federal University of Technology Parana, Brazil
Shi-Jay Chen	National United University, Taiwan
Shiuh-Jeng Wang	Central Police University, Taiwan
Izzettin Temiz	Gazi University, Turkey
Akira Asano	Hiroshima University, Japan
Andri Riid	Tallinn University of Technology, Estonia
Sylvain Piechowiak	University of Valenciennes, France
Wang Feng	Kunming University of Science and Technology, China
Ming-Wen Hu	Tamkang University, Taiwan
Kun-Huang Kuo	Chienkuo Technology University, Taiwan
Yi-Nung Chung	National Changhua University of Education, Taiwan
Jenn-Kaie Lain	National Yunlin University of Science and Technology, Taiwan
Tsung-Chih Lin	Feng-Chia University, Taiwan
Chyuan-Huei Thomas Yang	Hsuan Chuang University, Taiwan
Eiji Uchino	Yamaguchi University, Japan
Shu-Hua Hua	Jinwen University of Science and Technology, Taiwan
Rung-Ching Chen	Chaoyang University of Technology, Taiwan
Shie-Jue Lee	National Sun Yat-Sen University, Taiwan
Yong Zhang	Shenzhen University, China
Luciano Sanchez	Oviedo University, Spain
Donato Impedovo	Politecnico di Bari, Italy
Hsu-Yang Kung	National Pingtung University of Science and Technology, Taiwan
Tien-Tsai Huang	Lunghwa University of Science and Technology, Taiwan
Zhiyong Zhang	Henan University of Science and Technology, China
Yuh-Chung Lin	Tajen University, Taiwan
Guan-Hsiung Liaw	I-Shou University, Taiwan
Yuchi Ming	Huazhong University of Science and Technology, China
S.N. Omkar	Indian Institute of Science, Indian

Mohammed Al Rashidi	Public Authority for Applied Education and Training (PAAET), Kuwait
Jie Jing	Zhejiang University of Technology, China
Maurice Clerc	Independent Consultant, France
K.W. Wong	City University of Hong Kong, Hong Kong
S.N. Singh	Indian Institute of Technology Kanpur, Indian
Zhihua Cui	Taiyuan University of Science and Technology, China
Marco Mussetta	Politecnico Di Torino, Italy
Ling Wang	Tsinghua University, China
Shing Chiang Tan	Multimedia University, Malaysia
Jonathan Hoyin Chan	King Mongkut's University of Technology Thonburi, Thailand
Yin Chai Wang	University Malaysia Sarawak, Malaysia
Weng Kin Lai	MIMOS Berhad, Malaysia
G. Sainarayanan	ICT Academy of Tamil Nadu, India
Sheng-Yuan Yang	St. John's University, Taiwan
Mariacarla Staffa	Universita' degli Studi di Napoli "Federico II", Italy
Wu-Chih Hu	National Penghu University of Science and Technology, Taiwan
Deng-Yuan Huang	Dayeh University, Taiwan
Pei-Yin Chen	National Cheng Kung University, Taiwan
Tsung-Han Tsai	National Central University, Taiwan
Zhijian Wu	Wuhan University, China
Sanyou Zeng	China University of Geosciences (Wuhan), China
Lei Wang	Xi'an University of Technology, China
Qingzheng Xu	Xi'an Communication Institute, China
Mu-Song Chen	Da-Yeh University, Taiwan
Hsiang-Cheh Huang	National University of Kaohsiung, Taiwan
Philippe Fournier-Viger	University of Moncton, Canada

Contents

Part I
Data Mining Techniques and Its Applications

An Efficient Solution for Time-Bound Hierarchical Key Assignment Scheme

Jeng-Shyang Pan, Tsu-Yang Wu, Chien-Ming Chen and Eric Ke Wang

Abstract Time-bound hierarchical key assignment (TBHKA) scheme is a cryptographic method. It can assign encryption keys depending on time to a set of security classes in a partially ordered hierarchy. Only the authorized user can compute the encryption key to access the subscribing class (including lower down class) according to the hierarchy. In 2005, Yeh firstly proposed a RSA-based TBHKA scheme supporting discrete time period. However, it had been proved insecure against user colluding attacks. Up to now, there are less study for TBHKA scheme supporting discrete time period. In this paper, we propose a secure and efficient TBHKA scheme. Our scheme is based on pairing-based public key cryptosystem and supports discrete time period. The security analysis is demonstrated that our scheme is secure against outside adversary and malicious user. Finally, we make comparisons between recently proposed two TBHKA schemes and our scheme. It will show the advantages of our scheme.

Keywords Hierarchical key assignment · Time-bound · Bilinear pairing · Security

1 Introduction

The access control (AC) problem is to dominate who can access the sensitive resources in a system. According to users priority, users are organized in a hierarchy

J.-S. Pan
College of Information Science and Engineering, Fujian University of Technology, Fuzhou 350118, China
e-mail: jengshyangpan@gmail.com

T.-Y. Wu · C.-M. Chen · E.K. Wang
Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen 518055, China

T.-Y. Wu(✉) · C.-M. Chen · E.K. Wang
Shenzhen Key Laboratory of Internet Information Collaboration, Shenzhen 518055, China
e-mail: {wutsuyang,chienming.taiwan}@gmail.com, 962982698@qq.com

T.T. Zin et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 388,
DOI: 10.1007/978-3-319-23207-2_1

formed by several disjoint classes called security classes. These classes have different limitations on the resources. In other words, some users own more access rights than others. In the real world, the AC problem is applied to several applications such as hospital system, computer system, etc.. For example, in computer system, administrator has the high priority to access all files including sensitive files, but general users only access some common files. Up to now, several famous hierarchical key assignment schemes [1, 4, 10] had been published to solve the AC problem.

In some applications, time-bound property may be involved in the AC problem such as Pay-TV system. In Pay-TV system, subscriber desires to subscribe some channels for some certain time periods such as three month, half year, or one year. Hence, subscribers should be assigned different keys for each time period. If the time period expires, the subscriber should not derive any keys to access subscribed channels. Time-bound hierarchical key assignment scheme is a cryptographic method. It assigns several encryption keys to a set of security classes in a partially ordered hierarchy, where the keys are dependent on the time. Note that if two classes form a relation, the subscriber who is in the higher class can access the resources in the lower class, however not vice versa.

In 2002, Tzeng [12] proposed the first time-bound hierarchical key assignment (TBHKA) scheme by using Lucas function. However, Yi and Ye [18] showed that his scheme suffered from a user colluding attack. In 2004, Chien [8] proposed an efficient TBHKA scheme by using two hash values. Unfortunately, his scheme also suffered from a user colluding attack [17]. In 2005, Yeh [16] proposed an RSA-based hierarchical key assignment scheme. Yeh's scheme is the first TBHKA scheme supporting discrete time period. However, his scheme also suffered a user colluding attack [2]. In 2006, Ateniese et al. [3] defined a unconditionally secure and computationally secure setting for a TBHKA scheme. They also proposed a secure pairing-based TBHKA scheme supporting discrete time period. In the same year, Wang and Laih [13] proposed a TBHKA scheme by using merging. In 2012, Chen et al. [5] proposed a TBHKA scheme without tamper-resistant device. In the same year, Tseng et al. [11] proposed two time-bound key management schemes without hierarchy. The first scheme combines pairing-based public key cryptosystem and Lucas function [12] and is suitable for continuous time period. The second scheme fuses pairing-based and RSA public key cryptosystems and is suitable for discrete time period. Recently, many researchers focus on the study of TBHKA in cloud computing. In 2013, Chen et al. [6] proposed the first hierarchical access control scheme in cloud computing named *CloudHKA*. They used *CloudHKA* to encrypt outsourced data so that the resulted data are secure against honest but curious cloud server. In 2014, Wu et al. [15] extended Chen et al.'s scheme [5] to propose the first TBHKA scheme in cloud computing. Later on, He et al. [9] proposed an efficient solution for TBHKA in cloud environment.

Up to now, there are less study for TBHKA scheme supporting discrete time period. In this paper, we propose a secure and efficient TBHKA scheme. Our scheme is based on pairing-based public key cryptosystem and supports discrete time period. The security analysis is demonstrated that our scheme is secure against outside adversary and malicious user. Finally, we make comparisons between recently

proposed two TBHKA schemes [3, 16] and our scheme. We can find that our scheme is more efficient than Ateniese et al.'s scheme [3].

The rest of this paper is organize as follows: In Section 2, we introduce the related background knowledge for our TBHKA scheme. Our concrete scheme is proposed in Section 3. In Section 4, we demonstrate the security analysis of our scheme. The performance analysis and comparisons are made in Section 5 and the conclusions are draw Section 6.

2 Background Knowledge

In this section, we brief introduce the related background knowledge of our scheme including partially ordered hierarchy, bilinear pairings, and the discrete logarithm assumption in elliptic curve.

2.1 Partially Ordered Hierarchy

In a partially ordered hierarchy $(\mathfrak{C}, \preceq)$, there is a binary relation $\preceq$ partially orders the set of classes $\mathfrak{C}$. For any two classes $C_i, C_j \in \mathfrak{C}$, the notation $C_j \preceq C_i$ means that the priority of C_i is higher then C_j. Hence, a user in C_i can access C_j and the opposite is forbidden. Obviously, $C_i \preceq C_i$ for any $C_i \in \mathfrak{C}$. In addition, the partially ordered hierarchy $(\mathfrak{C}, \preceq)$ can be represented by a directional graph, where each class corresponds to a vertex in the graph and there exists an edge from class C_i to C_j if and only if $C_j \preceq C_i$. For the detailed descriptions about partially ordered hierarchy, readers can refer to [3, 12].

2.2 Bilinear Pairings

Let G_1 and G_2 be two groups with a same large prime order q, where G_1 is an additive cyclic group of an elliptic curve $E(F_p)$ over a finite field F_p and G_2 is a multiplicative cyclic group of F_p. A bilinear pairing e is a map defined by $e : G_1 \times G_1 \rightarrow G_2$ and satisfies the following three properties:

1. Bilinear: For all $P, Q \in G_1, a, b \in \mathbb{Z}_q^*$, we have $e(aP, bQ) = e(P, Q)^{ab}$.
2. Non-degenerate: For all $P \in G_1$, there exists $Q \in G_1$ such that $e(P, Q) = 1_{G_2}$.
3. Computable: For all $P, Q \in G_1$, there exists an efficient algorithm to compute $e(P, Q)$.

The detailed descriptions for bilinear pairings can be referred to [14].

2.3 Discrete Logarithm (DL) Problem in Elliptic Curve

- *Discrete logarithm (DL) problem*: Given $P, aP \in G_1$ for $a \in Z_q^*$, the DL problem is to compute the value a.
- *DL assumption*: No probabilistic polynomial time (PPT) algorithm can solve the DL problem.

3 Proposed Scheme

In this section, we propose an efficient time-bound hierarchical key assignment scheme. In the proposed scheme, we assume that each user can access some resources in a set T_i of discrete time intervals, for example {January, March, June}. Without loss of generality, we assume that these resources are stored in a set of classes $\mathfrak{C} = \{C_1, C_2, \ldots, C_n\}$. Note that the n classes form a directional graph with the relation $\preceq$ mentioned in Subsection 2.1. Meanwhile, we define the maximal system life time as $T = \{1, 2, \ldots, z\}$, ie. $T_i \subset T$. The proposed scheme consists of following three phases: *System setup*, *User subscribing*, and *Encryption key derivation* phases.

3.1 System Setup Phase

The system vender (SV) constructs a set of classes $\mathfrak{C}$ and deploys the resources into n classes. It means that a directional graph $(\mathfrak{C}, \preceq)$ is produced. Then, the SV generates the needed keys and parameters as follows.

1. The SV selects a elliptic curve E over a finite field F_p and defines a bilinear map $e : G_1 \times G_1 \rightarrow G_2$ mentioned in Subsection 2.2. A generator $P \in G_1$ is generated and then a public value $P_{pub} = s \cdot P$ is computed, where $s \in \mathbb{Z}_q^*$ is a secret value kept by the SV.
2. For each class C_i, the system vender selects a secret value $a_i \in \mathbb{Z}_q^*$ kept by the SV. Then, for each time period $t \in \{1, 2, \ldots, z\}$, the system vender chooses other secret values $b_t \in \mathbb{Z}_q^*$, $t = 1, 2, \ldots, z$, kept by the SV.
3. For each $C_j \preceq C_i$, ie. user in C_i can access C_j (there exists a edge from C_i to C_j), the system vender computes the public value $P_{i,j} = (a_j/a_i) \cdot P_{pub}$.
4. In order to protect the resource of class C_i in time period $t \in \{1, 2, \ldots, z\}$, the system vender compute an encryption key $K_{i,t} = e(a_i P, P_{pub})^{b_t}$.

Finally, the SV publishes public parameters $\{e, G_1, G_2, q, P, P_{pub}, P_{i,j}\}$.

3.2 User Subscribing Phase

When a user subscribes class C_i in time period $T_i \subset \{1, 2, \ldots, z\}$, the system vender computes a key $a_i b_t P$ for $t \in T_i$ and sends it to the user. Note that $a_i b_t P$ can be used to derive the encryption key $K_{i,t}$.

3.3 Encryption Key Derivation Phase

For any user in class C_i with her/his subscribing time period T_i, she/he can compute the decryption key $K_{j,t}$ of class C_j if and only if $C_j \preceq C_i$ and $t \in T_i$. The key derivation is shown as follows:

$$K_{j,t} = e(P_{i,j}, a_i b_t P) = e((a_j/a_i) \cdot P_{pub}, a_i b_t P) = e(a_j P, P_{pub})^{b_t}.$$

4 Security Analysis

In this section, we demonstrate the security of our scheme. Since the encryption key $K_{i,t}$ is computed by the public value $P_{i,j}$ and key $a_i b_t P$, the security of $K_{i,t}$ relies on (1) the security of $P_{i,j}$ and (2) the security of $a_i b_t P$.

Lemma 1. *Under the discrete logarithm assumption in elliptic curve, any outside adversary and inside user cannot obtain the secret value a_i and a_j from $P_{i,j} = (a_j/a_i) \cdot P_{pub}$ for all i and j.*

Proof. By the discrete logarithm assumption, no probabilistic polynomial time algorithm can be used to compute the secret value a_i and a_j from $P_{i,j}$. In other words, given a public value $P_{i,j} = (a_j/a_i) \cdot P_{pub}$, anyone cannot obtain the secret values a_i and a_j from $P_{i,j}$ for some i and j.

Lemma 2. *Under the discrete logarithm assumption in elliptic curve, any outside adversary and malicious user cannot compute an unauthorized key $a_i b_t P$ to access some class C_i in some time period t.*

Proof. Here, we consider the two cases:

(1) For the outside adversary, he wants to compute an unauthorized key $a_i b_t P$ for some i and t. By Lemma 1, any outside adversary cannot obtain the secret value a_i from public value $P_{i,j}$. Meanwhile, the value b_t is kept by the system vender secretly. Thus, this case is impossible.
(2) For the malicious user, he wants to access some unsubscribed class or access some class in an unsubscribed time period. By the same reason from Case (1), it is impossible.

By Lemmas (1) and (2), we have proven the proposed scheme is a secure time-bound hierarchical key assignment scheme.

Theorem 1. *Under the discrete logarithm assumption in elliptic curve, the proposed scheme is a secure time-bound hierarchical key assignment scheme.*

5 Performance Analysis and Comparisons

In this section, we make the performance analyze and comparisons between two time-bound hierarchical key assignment schemes supporting discrete time periods (Ateniese et al.'s scheme [3] and Yeh's scheme [16]), and our scheme. In order to convenience to evaluate the computational cost, we first define the following notations:

- TG_e: The time of executing a bilinear pairing operation, $e : G_1 \times G_1 \rightarrow G_2$.
- TG_{mul}: The time of executing a scalar multiplication operation of point in G_1.
- T_{exp}: The time of executing a modular exponentiation operation.
- T_{inv}: The time of executing a modular multiplicative inverse operation.

In the system setup phase, it requires $TG_e + T_{exp}$ to generate the encryption key $K_{i,t}$ for class C_i in time period t. It requires TG_{mul} to compute a user key $a_i b_t P$ in the user subscribing phase. In the encryption key derivation phase, it requires TG_e to derive the encryption key $K_{i,t}$. Finally, to compute the public value $P_{i,j}$ requires $TG_{mul} + T_{inv}$ computational cost.

In the following Table 1, we demonstrate the comparisons between Ateniese et al.'s scheme [3], Yeh's scheme [16], and our scheme in terms of the key construction and the computational cost of each phase. It is easy to see that Yeh's scheme [16] is based on the RSA public key cryptosystem and Ateniese et al.'s scheme [3] and our scheme are based on the pairing-based public key cryptosystem. However, Yeh's scheme [16] suffered user colluding attack mentioned by Ateniese et al. [3].

On the other hand, the performance of our scheme and Ateniese et al.'s scheme [3] are the same except the user key generation and the public value generation. In [7], the simulation results show that one modular exponentiation operation is about 2.5 times scalar multiplication operation of point in G_1, ie. $T_{exp} \approx 2.5TG_{mul}$. Hence, our scheme is more efficient than Ateniese et al.'s scheme [3]. In addition, both schemes are provably secure.

Table 1 The comparison between the recent proposed time-bound hierarchical key assignment schemes and our scheme

	Yeh's scheme [16]	Ateniese et al.'s scheme [3]	Our scheme
Public key construction	RSA	Pairing based	Pairing based
Encryption key generation	$2T_{exp}$	$TG_e + TG_{mul} + T_{exp}$	$TG_e + TG_{mul} + T_{exp}$
User key generation	T_{exp}	$T_{exp} + T_{mul}$	$TG_{mul} + T_{mul}$
Encryption key derivation	T_{exp}	TG_e	TG_e
Public value generation	$\times$	$T_{exp} + T_{inv}$	$TG_{mul} + T_{inv}$
Security	Colluding attacks [2]	Provably secure	Provably secure

6 Conclusions

In this paper, we have proposed a time-bound hierarchical key assignment (TBHKA) scheme. The security analysis is demonstrated that our scheme is provably secure against outside adversary and malicious user. With comparing the recently proposed two TBHKA schemes supporting discrete time period, our scheme is very efficient.

Acknowledgments This work is supported by Shenzhen Strategic Emerging Industries Program of China (No. ZDSY20120613125016389).

References

1. Akl, S.G., Taylor, P.D.: Cryptographic solution to a problem of access control in a hierarchy. ACM Transactions on Computer Systems (TOCS) **1**(3), 239–248 (1983)
2. Ateniese, G., De Santis, A., Ferrara, A.L., Masucci, B.: Provably-secure time-bound hierarchical key assignment schemes. In: Proceedings of the 13th ACM Conference on Computer and Communications security, pp. 288–297. ACM (2006)
3. Ateniese, G., De Santis, A., Ferrara, A.L., Masucci, B.: Provably-secure time-bound hierarchical key assignment schemes. Journal of Cryptology **25**(2), 243–270 (2012)
4. Blanton, M., Fazio, N., Frikken, K.B.: Dynamic and efficient key management for access hierarchies. In: Proceedings of the ACM Conference on Computer and Communications Security (2005)
5. Chen, C.M., Wu, T.Y., He, B.Z., Sun, H.M.: An efficient time-bound hierarchical key management scheme without tamper-resistant devices. In: 2012 International Conference on Computing, Measurement, Control and Sensor Network (CMCSN), pp. 285–288. IEEE (2012)
6. Chen, Y.-R., Chu, C.-K., Tzeng, W.-G., Zhou, J.: Cloudhka: a cryptographic approach for hierarchical access control in cloud computing. In: Jacobson, M., Locasto, M., Mohassel, P., Safavi-Naini, R. (eds.) ACNS 2013. LNCS, vol. 7954, pp. 37–52. Springer, Heidelberg (2013)
7. Cheng, Z.: Implementing pairing-based cryptosystems in usb tokens. IACR Cryptology ePrint Achieve **2014**(71) (2014)
8. Chien, H.Y.: Efficient time-bound hierarchical key assignment scheme. IEEE Transactions on Knowledge and Data Engineering **16**(10), 1301–1304 (2004)
9. He, B.Z., Chen, C.M., Wu, T.Y., Sun, H.M.: An efficient solution for hierarchical access control problem in cloud environment. Mathematical Problems in Engineering 2014, Article ID 569397, p. 8 (2014)
10. Jiang, T., Zheng, S., Liu, B.: Key distribution based on hierarchical access control for conditional access system in dtv broadcast. IEEE Transactions on Consumer Electronics **50**(1), 225–230 (2004)
11. Tseng, Y.M., Yu, C.H., Wu, T.Y.: Towards scalable key management for secure multicast communication. Information Technology And Control **41**(2), 173–182 (2012)
12. Tzeng, W.G.: A time-bound cryptographic key assignment scheme for access control in a hierarchy. IEEE Transactions on Knowledge and Data Engineering **14**(1), 182–188 (2002)
13. Wang, S.Y., Laih, C.S.: Merging: an efficient solution for a time-bound hierarchical key assignment scheme. IEEE Transactions on Dependable and Secure Computing **3**(1), 91–100 (2006)
14. Wu, T.Y., Tseng, Y.M.: An id-based mutual authentication and key exchange protocol for low-power mobile devices. The Computer Journal **53**(7), 1062–1070 (2010)
15. Wu, T.-Y., Zhou, C., Wang, E.K., Pan, J.-S., Chen, C.-M.: Towards time-bound hierarchical key management in cloud computing. In: Pan, J.-S., Snasel, V., Corchado, E.S., Abraham, A., Wang, S.-L. (eds.) Intelligent Data Analysis and Its Applications, Volume I. AISC, vol. 297, pp. 31–38. Springer, Heidelberg (2014)
16. Yeh, J.H.: A secure time-bound hierarchical key assignment scheme based on rsa public key cryptosystem. Information Processing Letters **105**(4), 117–120 (2008)
17. Yi, X.: Security of chien's efficient time-bound hierarchical key assignment scheme. IEEE Transactions on Knowledge and Data Engineering **17**(9), 1298–1299 (2005)
18. Yi, X., Ye, Y.: Security of tzeng's time-bound key assignment scheme for access control in a hierarchy. IEEE Transactions on Knowledge and Data Engineering **15**(4), 1054–1055 (2003)

Quaternion Principal Component Analysis for Multi-modal Fusion

Meng Chen, Chenxia Wang, Xiao Meng and Zhifang Wang

Abstract This paper proposes a multi-modal fusion method that based on quaternion, and principal component analysis (PCA) in quaternion field is involved in our algorithm. We can fuse four different features into quaternion and complete the recognition process in quaternion field. This algorithm reduces the equal error rate (EER) while fusing more kinds of features. Our experiments that fuses three kinds of modalities and four different features with two kinds of modalities respectively show a observably improvement on recognition rate with the proposed algorithm.

Keywords Quaternion field · Multi-modal fusion · PCA · KPCA

1 Introduction

Biometric identification technology achieve the individual identification by physiological(fingerprint, iris, face, etc.) and behavioral(gait, signature, voice, etc.) characteristics with inherent human body.[1] Compare with the traditional identity authentication, biological feature is universal, stable and sole. So the biometric identification technology has better safety, reliability and validity. It can not only achieve the authentication function, but also complete the binding of user ID and user's biological feature. With the development of computer technology and the improvement of the demand of social public security and personal information security, biometric identification technology has been widely used

M. Chen · X. Meng · Z. Wang(✉)
Department of Electronic Engineering, Heilongjiang University, Harbin, Heilongjiang, China
e-mail: {181363682,354911741}@qq.com, xiaofang_hq@126.com

C. Wang
Zhengzhou Kindergartens Teachers' College, Zhenzhou, Henan, China
e-mail: 553792191@qq.com

T.T. Zin et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 388,
DOI: 10.1007/978-3-319-23207-2_2

and developed in various fields. However, with the use of biometric identification, some problems of the traditional unimodal biometric identification have been appeared. It is mainly reflects in recognition performance and system security. Although the recognition rates of sundry unimodal biometric identification technology have been higher, it is not enough for some complex environments. For example the airport and station where has a high flow of people. Meanwhile, the progress of forgery technology makes the unimodal biometric identification be easy to cheat by false features. The multi-modal biological feature fusion technology can overcome these above defects. So it has a speedy development in recent years.

It is start-up phase of multi-modal biological feature fusion from 1995 to 2000. The concept of multi-modal is came up by Bigun and Duc [2,3] in 1997. Their method fused voice and face features by supervised learning and Bayesian theory with a high accuracy. With the development of multi-modal fusion technology, there are many feasible fusion algorithm that are came up in recent years. Ling Lin [4] proposed a fusion method that can fuse face and palm print features in the feature level in 2011. She extracted these biometrics features with Gabor wavelet and two dimensional principal component analysis techniques. And she fused biometric features of two modalities with a new weighting strategy. There are also face-ear fusion [5], palm print-hand vein fusion [6], etc. However, most of them can just fuse two single-feature modalities or one modality with two kinds of feature. Looking up the existing algorithms, Serial rule [7] and weighted sum rule [8] are two kinds of famous fusion method in the feature level. They can both improve the recognition rate comparing with single modality. But it is not good enough.

In this paper, we will propose a quaternion based fusion algorithm that can fuse four different features at most. And the recognition rate of the proposed algorithm is better than single feature, serial rule and weighted sum rule. The content of this paper can be summarized as follows: Firstly, introduce the feature extraction method of different kinds of modalities. Secondly, give the related knowledge of quaternion. Then, propose the fusion method of these extracted features and expand PCA to the quaternion field. Finally, give the result of our experiments.

2 Propose Algorithm

2.1 Multi-feature Extraction

We fuse different biological modality in the feature level. At the beginning of our algorithm, it is needed to extract features. In this stage, we involve the PCA [9] and the Kernel Principal Component Analysis (KPCA) [10]. PCA gives linear features and KPCA gives non-linear features. When there are two modalities we need to fuse, we respectively extract the one with PCA and KPCA and then the other one. We can get four kinds of features with this method. When there are three modalities, we extract all of them with PCA and get three kinds of features.

2.2 Quaternion Based Multi-modal Fusion

2.2.1 Quaternion and Quaternion Matrix

$q = a + bi + cj + dk$ is named quaternion while $a, b, c, d \in R$ and i, j, k meet $i^2 = j^2 = k^2 = -1$, $ij = -ji = k$, $jk = -kj = i$, $ki = -ik = j$. Quaternion q is made up of the real part and imaginary part, respectively are $a$and $bi + cj + dk$. According to imaginary multiplication rule, we can also rewrite quaternion q as $q = (a + ib) + (c + di)j$. If $q_1 = a_1 + b_1 i + c_1 j + d_1 k$, $q_2 = a_2 + b_2 i + c_2 j + d_2 k$, then we can define the basal operation principles of quaternion by the following equation.

Equality: $q_1 = q_2 \Leftrightarrow a_1 = a_2, b_1 = b_2, c_1 = c_2, d_1 = d_2.$

Addition and subtraction: $q_1 \pm q_2 = (a_1 \pm a_2) + (b_1 \pm b_2)i + (c_1 \pm c_2)j + (d_1 \pm d_2)k.$

Multiplication: $q_1 \cdot q_2 = (a_1 a_2 - b_1 b_2 - c_1 c_2 - d_1 d_2) + (a_1 b_2 + b_1 a_2 + c_1 d_2 - c_2)i + (a_1 c_2 + a_2 c_1 + b_2 d_1 - d_2 b_1)j + (a_1 d_2 + d_1 a_2 + c_2 - c_1 b_2)k.$

Square: $q^2 = (a^2 - b^2 - c^2 - d^2) + 2abi + 2acj + 2adk.$

Conjugate: $\bar{q} = a - bi - cj - dk.$

Rules: $N(q) = q\bar{q} = \bar{q}q = a^2 + b^2 + c^2 + d^2 \geq 0.$

Modulus: $|q| = \sqrt{N(q)} = \sqrt{a^2 + b^2 + c^2 + d^2}.$

If Q is the set of quaternion and matrix $A \in Q^{n \times n}$, then the following concepts and propositions are involved in the algorithm we proposed.

1) If $A^H = A$, then A is named self-conjugate quaternion matrix. $SC_n(Q)$ called the set of $n$order self-conjugate quaternion matrix;

2) If exist $\lambda \in Q$ and $0 \neq a \in Q^{n \times 1}$ make $Aa = a\lambda$(or $Aa = \lambda a$), then call λ the right(or left) eigenvalue of A and a is the eigenvector belong to right(or left) eigenvalue of A. Then name λ the eigenvalue of A, if λ is both the right and left eigenvalue of A;

3) If $U \in Q^{n \times n}$ and $UU^H = U^H U = I$, then U is named extended unitary matrix.

4) Quaternion matrix can expressed as $= A_1 + A_2 j$. That is the sum of two plural matrices we mentioned in the first paragraph. Then the induced matrix of quaternion matrix A can be write as plural matrix $A^\sigma = \begin{pmatrix} A_1 & -A_2 \\ \overline{A_2} & \overline{A_1} \end{pmatrix}$;

5) If A is self-conjugate quaternion matrix, then A^σ is Hermite matrix;

6) Quaternion matrix A and its induced matrix A^σ has the same eigenvalues;

7) If quaternion matrix A has a induced matrix A^{σ}, and the eigenvalue of A^{σ} is λ, the corresponding eigenvector is $\begin{pmatrix}\alpha_1\\ \alpha_2\end{pmatrix}$, then the eigenvector of λ that is the eigenvalue of quaternion matrix A is $\alpha_1 + \overline{\alpha_2}j$;

If quaternion $q = a + bi + cj + dk$, then it has the trigonometric expression as follow.

$$q = |q|(cos\,\theta + Isin\,\theta)$$

$$I = \frac{1}{\sqrt{b^2 + c^2 + d^2}}(bi + cj + dk)$$

$$\theta = arctan\frac{\sqrt{b^2 + c^2 + d^2}}{a}$$

Then the formula of quaternion n-th root expresses as

$$\sqrt[n]{q} = \sqrt[n]{|q|}\left(cos\frac{\theta + 2k\pi}{n} + I\,sin\frac{\theta + 2k\pi}{n}\right), k = 0,1,2,\cdots,n$$

If Q is a quaternion vector, the Euclidean norm of Q expresses as

$$\|\mathrm{Q}\|_2 = (\mathrm{Q},\mathrm{Q})^{\frac{1}{2}} = \left(\sum_{\mathrm{i}-1}^{\mathrm{n}} \mathrm{q}_{\mathrm{i}}^2\right)^{\frac{1}{2}}$$

2.2.2 Orthogonal Eigenvectors

The associative law and the commutative law of quaternion addition and the associative law of quaternion multiplication are easy to verify. But the commutative law of multiplication is not applicatory in the quaternion field. The operation of quaternion matrix is more complex than real or plural matrix according to the above reason. If we want to expand the PCA to the quaternion field, it is necessary to figure out the orthogonal eigenvectors of quaternion matrix. In theory, we can figure out the eigenvectors with the method mentioned in the above proposition 7. However, the eigenvectors we get with the proposition 7 method is uncertain to be mutually orthogonal in the process of practice programme. In view of this, we reference a reasonable way proposed by LANG Fang-Nian [11] to solve this problem.

A random $n \times n$ self-conjugate quaternion matrix A (the train sample scatter matrix in this multi-modal fusion recognition application is a self-conjugate quaternion matrix), its induced matrix is A^{σ}. And $A^{\sigma}X = \lambda X$ is the feature equation of A^{σ}, that is $(A^{\sigma} - \lambda I)X = 0$. Let $\lambda_i \in R$ be the eigenvalues and the corresponding eigenvectors of A^{σ} are $X_i, i = 1,2,\cdots,n$. Build the following expressions with different λ_i, λ_j.

$$\mathrm{I} - (\mathrm{A}^{\sigma} - \lambda_{\mathrm{i}}\mathrm{I})^{+}(\mathrm{A}^{\sigma} - \lambda_{\mathrm{i}}\mathrm{I})$$
$$\mathrm{I} - \left(\mathrm{A}^{\sigma} - \lambda_{\mathrm{j}}\mathrm{I}\right)^{+}\left(\mathrm{A}^{\sigma} - \lambda_{\mathrm{j}}\mathrm{I}\right)$$

If $\begin{pmatrix}\alpha_1\\ \alpha_2\end{pmatrix}$ and $\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix}$ respectively are the first column vector of the above two expressions, and α_1, α_2, β_1, β_2 have the same dimensions. Then λ_i, λ_j are the eigenvalues of quaternion matrix A, and $\alpha_1 + \overline{\alpha_2}j$, $\beta_1 + \overline{\beta_2}j$ are the corresponding eigenvectors. In addition, they are proved to be orthogonal.

2.2.3 Expand to Quaternion Field

Quaternion q is consist of four parts which are one real part a and three imaginary parts $bi + cj + dk$. Let these features extracted before build the four or three parts of quaternions. For example, if there are feature A, feature B, feature C and feature D. We can build a quaternion q like $q = A + Bi + Cj + D$. And if there are feature A, B and C, we can build it like $q = 0 + Ai + Bj + C$. In this way, We can get an quaternion matrix that consist of these four or three different features. The training and testing sample set are built in this method. We finish the algorithm and deal with the sample set in the quaternion field with PCA. So, it is necessary to expand the general PCA to the quaternion field.

It's easy to expand the PCA to quaternion according to the method introduced in the 2.2.2. If $\{X \parallel x_i, i = 1,2,\cdots,m\}$ and $x_i = (\alpha_{i1}\ \alpha_{i2}\ \cdots\ \alpha_{in}), \alpha_{ik} \in Q$ is the normalized standard training sample set. Then the generating matrix should be

$$\sum = \frac{1}{M}\sum_{i=1}^{M}(x_i - \mu)(x_i - \mu)^H$$

x_i is the ith training sample vector, μ is the average vector of sample set and M is the total of training sample. Figure out the orthogonal eigenvector set of the generating matrix with the method proposed in the 2.2.2, and choose same eigenvectors which correspond these bigger eigenvalues as principal vectors, then the PCA can be applied to quaternion field with general definition in the field of real number. It's worth noting that the generating matrix structured with above method is a self-conjugate quaternion matrix.

3 Results and Analysis

We performed two experiments about our algorithm. Figure 1 gives the process of one experiment. It involves four features with two kinds of modalities which are face and palm print. We extract facial feature with PCA and KPCA, extract palmprint feature with PCA and KPCA. Then fuse them like $q = A + Bi + Cj + D$ that we mentioned before. Figure 2 gives the process of another experiment that involves three features which respectively are face, palm print and signature. We extract these features with PCA and fuse them like $q = 0 + Ai + Bj + C$. Then the fusion quaternion feature are dealt with quaternion PCA.

In our experiments, modal images are provided by Yale face database, PolyU palm print database and ATVS-Synthetic Signature database. Yale face database is made by Yale university center for computational vision and control, 15

volunteers, 165 pictures with different illumination, expression and posture are involved in it. PolyU palm print database is build by The Hong Kong Polytechnic University and involves 100 volunteers with 6 pictures each. And ATVS-Synthetic Signature database contains 25 signatures of 350 users.

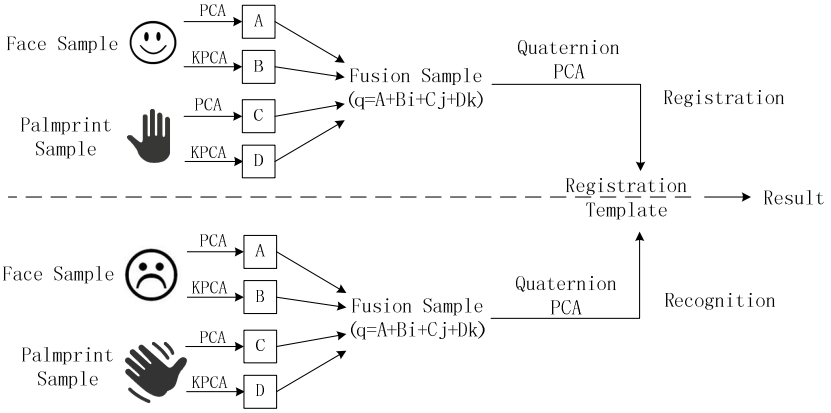


Fig. 1 The process of the proposed algorithm(four features of two modalities).

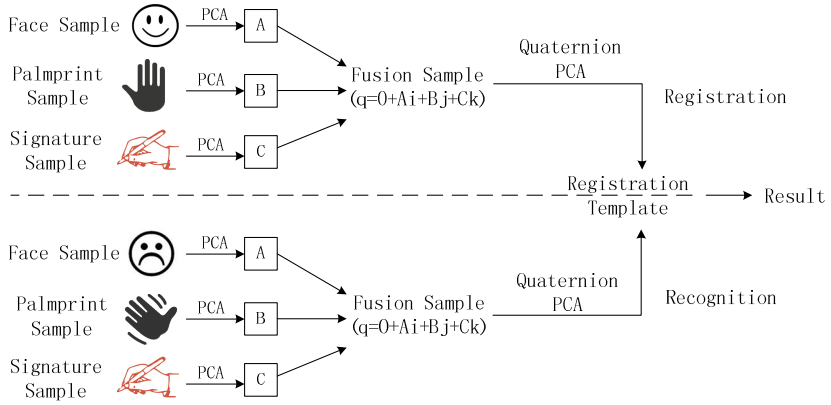


Fig. 2 The process of the proposed algorithm(three modalities).

False match rate(FMR) and False non-match rate(FNMR) are the major two parameters of recognition algorithm performance evaluation. The equal error rate(EER) can measure the overall performance of an algorithm. It unifies FMR and FNMR at the same time. FMR increases with the increase of the threshold and FNMR decreases with the increase of the threshold. EER refers to the value of the intersection of FNMR and FMR in the same coordinate. For a high-performance algorithm, there is a smaller value of EER. The DET curve is similar to EER, the x and y axis of it respectively are FMR and FNMR. Lower DET curve represents higher performance of algorithm.

Table 1 EER of two-modal(four features) fusion

Feature	EER(%)
Face with PCA	13.3
Face with KPCA	15.8
Palmprint with PCA	17.7
Palmprint with KPCA	19.7
Weighted sum rule	13.3
Serial rule	11.1
Proposed algorithm	6.6

Table 2 EER of three-modal fusion

Feature	EER(%)
Face	13.3
Palm print	17.7
Signature	6.1
Weighted sum rule	4.4
Serial rule	4.4
Proposed algorithm	2.2

Table1 gives the data of EER of four features with two kinds of modalities. Table 2 gives the data of EER of three modalities. And both of them involve the EER of weighted sum rule and serial rule. From these tables, we can clearly see that the EER of algorithm we proposed is lower than any single features and the other two fusion method. Figure 3 and figure 4 are the DET curves of these algorithms and the proposed algorithm is still the lowest of them. According to these experiments we did, it is obviously that our algorithm has a better performance than the others we mentioned. At the same time, it can fuse more features than most existing fusion method.

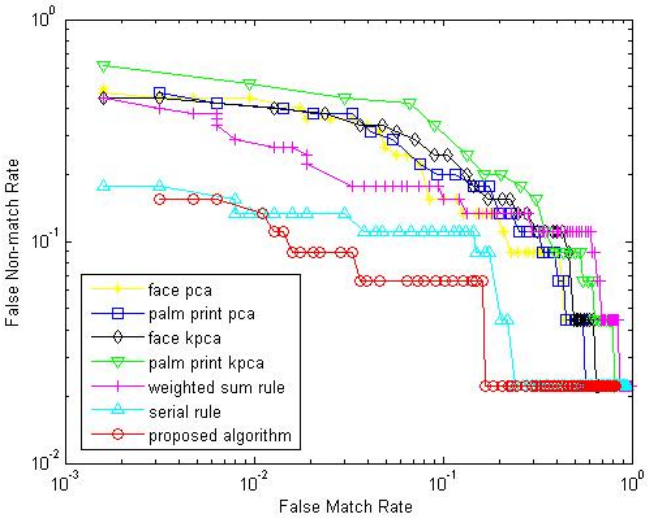


Fig. 3 DET of two-modal(four features) fusion

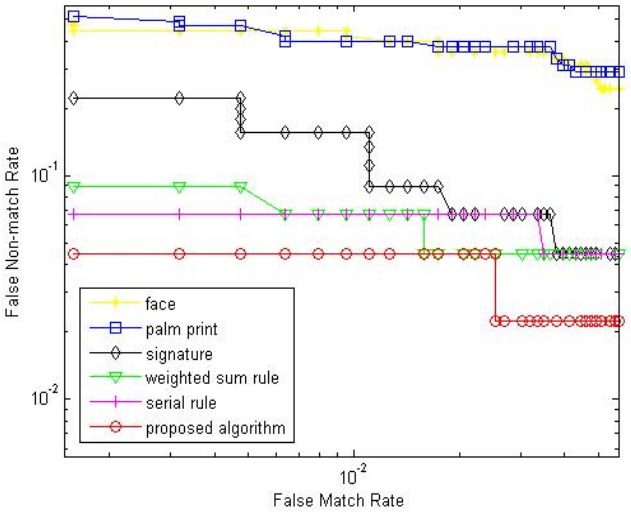


Fig. 4 DET of three-modal fusion

4 Conclusions

We proposed a new quaternion based multi-modal fusion algorithm in this paper. We fused different features into quaternion and generalized the general PCA to the quaternion field. Then achieved the identity recognition in the quaternion field. Comparing with most of these existing algorithms, we enhance the number of

feature from two to four. And our experiments showed the recognition rate have greatly improved than single feature and other two fusion method.

Acknowledgments. This work is supported by National Natural Science Foundation of China (no.61201399), China Postdoctoral Science Foundation (no.2012M511003), Project of Science and Technology of Heilongjiang Provincial Education Department (no.12521418), Youth Foundation of Heilongjiang University (no.201026), and Startup Fund for Doctor of Heilongjiang University.

References

1. Jain, A.K.: Biometric Recognition: Q&A. Nature, 449 (2007)
2. Bigün, E., Bigün, J., Duc, B., et al.: Expert conciliation for multimodal person authentication systems using bayesian statistics. In: Bigün, J., Borgefors, G., Chollet, G. (eds.) AVBPA 1997. LNCS, vol. 1206, pp. 291–300. Springer, Heidelberg (1997)
3. Duc, B., Bigün, E.S., Bigün, J., et al.: Fusion of Audio and Video Information for Multimodal Person Authentication. Pattern Recognition Letters **18**(9), 835–843 (1997)
4. Ling Lin, X., Zhou, M., et al.: Research on multimodal biometric authentication using feature level fusion. Computer Engineering and Design **35**(8), 2849–2852 (2011)
5. Yuan, L., Mu, Z.C., Zeng, H.: Multimodal recognition using face and ear. Journal of University of Science and Technology Beijing **29**(2), 191–193 (2007)
6. Ma, H.: Research of recognition method based on the feature layer fusion of palmprint and hand vein. Tianjin University of Technology (2014)
7. Rattani, A., Kisku, D.R., Bicego, M., et al.: Robust Feature-level Multibiometric Classification. In: Proceedings of Biometric Consortium Conference, pp. 1–6. IEEE, Baltimore, (2006)
8. Yao, Y., Jing, X., Wong, H.: Face and Palmprint Feature Level Fusion for Single Sample Biometrics Recognition. Neurocomputing Letters **70**(7–9), 1582–1586 (2007)
9. Jolliffe, I.T.: Principal Component Analysis. Springer Verlag, New York (1986)
10. Scholkopf, B., Smola, A.J., Muller, K.R.: Nonlinear Component Analysis as a Kernel Eigenvalue Problem. Nerual Computation **10**(5), 1299–1319 (1998)
11. Lang, F.N., et al.: Obtain Method of Quaternion Matrix Orthogonal Eigenvector Set and Its Application in Color Face Recognition. Acta Automatica Sinica **34**(2), 121–129 (2008)

A Novel Load Balance Algorithm for Cloud Computing

Linlin Tang, Jeng-Shyang Pan, Yuanyuan Hu, Pingfei Ren, Yu Tian and Hongnan Zhao

Abstract A good scheduling algorithm is a key for load balance system, in which system's load meets users' requirement. Here, a new load balance algorithm based on swarm intelligence is proposed which can enhance the production of the systems while schedule tasks to VMs properly. Here tasks completion time is compared with some other classical algorithms. The result shows that the proposed algorithm could meet users' requirement and get resource utilization higher. The algorithm is better for network of a large area which is simulated by CloudSim.

Keywords Composite sequence · Power spectrum · Direct sequence spread spectrum · Interference avoidance

1 Introduction

Cloud computing has recently emerged as a new paradigm of hosting and delivering services over the Internet [1][2]. That is a new style in which we won't compute on local computers, but on centralized facilities operated by third-party compute and storage utilities. The major computing is transferred from "endpoint" to "cloud"[3]. Since 2007, at the time when the concept of cloud computing was

L. Tang · J.-S. Pan(✉) · P. Ren · Y. Tian · H. Zhao
Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, China
e-mail: hittang@126.com, Jengshyangpan@gmail.com

J.-S. Pan
Fujian University of Technology, Fuzhou, China
e-mail: Jengshyangpan@gmail.com

Y. Hu
College of Information Engineering, Shenzhen University, Shenzhen, China

T.T. Zin et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 388,
DOI: 10.1007/978-3-319-23207-2_3

put forward, more and more corporations have set foot in this area, such as Microsoft, Google, Amazon and so forth. They provide cloud computing products themselves respectively. It is a trend of Internet development to the future. As one of the most important aspects of cloud computing, load balance has earned more and more people's attention these years. There are many different definitions for cloud computing, one of them can be described as below.

"A large-scale distributed computing paradigm that is driven by economies of scale, in which a pool of abstracted virtualized, dynamically-scalable, managed computing power, storage, platform and services are delivered on demand to external customers over the Internet [4]."

In fact, the aim of cloud computing is to outsource computing infrastructure to third parties. According to different types of service, three provision models where Clouds are used can be given: Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS). Virtualization, parallel computing, grid computing and distributed computing can be seen as the important techniques to cloud computing [5]. The basic structure diagram can be shown in the following Figure 1 [6].

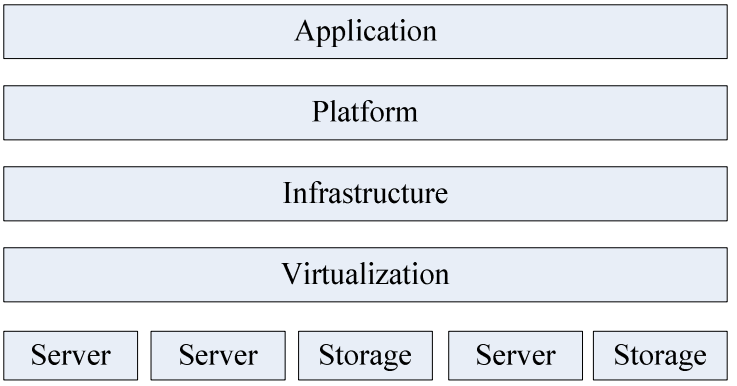


Fig. 1 The cloud computing scenarios diagram

Load balance algorithm, which is an important part of cloud computing, is the topic of this paper. It belongs to combinatorial optimization problems. It is also a kind of task scheduling problem. Here, an algorithm based on the Artificial Bee Colony is proposed and experiments is operated on Cloudsim [7][8], a cloud simulation toolkit, are used to show its efficiency.

Worker bees occupy the biggest amount in a bee swarm and they are also the most diligent ones in it. They work together to collect honey, product honey, investigate environment, guard the swarm and so on. In our algorithm, we simulated the behavior of worker bees based on the action they present at the time when they work on collecting honey. When collecting honey, based on different responsibilities, worker bees could be classified into scout bees, employed bees and onlooker bees.

When collecting honey, firstly, based on the multitude of nectar, the scout bees look for the flowers that are blooming which are the targets to collect honey. Secondly, after the flowers chosen, scout will dance to send messages to their buddies and tell them where the source is. Scout are then changed to employed bee which stays on the spot waiting for buddies. Due to rigorous division of labor and cooperation with each other, worker bees own an efficient way to collect honey.

From the above, It is obviously that the method bees collecting honey used is much better than some other random ones. In this paper, we originated using the artificial bee colony algorithm to solve the load balance problem exists in cloud computing. This is a new method by simulating the behavior of the worker bee to balance the load of the cloud computing system.

Additionally, because of the lack of hardware compared to the requirement in some cases, resource consumption may become of vital importance. Here the proposed Artificial Bee Colony algorithm could even consume little more than the Roud Robin one, thus having enough reason to choose it.

The content of this work is organized as follows: Related works are given in section 2, our proposed algorithm is shown in section 3. Experimental results and analysis are proposed in section 4. A conclusion for the whole and a look into the future are described in section 5. Acknowledgment is shown in section 6.

2 Related Works

Actually, more and more intelligent algorithms are also introduced to the cloud computing research area. Several important events are shown as below.

(1) M. Maheswaran put forwards dynamic heuristic allocation strategy in the task scheduling of distributed systems in the Hadoop [10].

(2)Carretero solved task scheduling problem in the network by using a genetic algorithm [11], aiming at getting the minimum flow time and the shortest job completing time.

(3)Jijian Xiong put forward the dynamic model of a heterogeneous environment [12], it solves resource assignment problems in a dynamic hetero generous environment through the interaction of swarm and environment.

(4)Kazem proposed an improved simulated annealing algorithm to solve the task scheduling problem in the network environment [13].

(5)Chen Yulanput forward a constraint algorithm based on service quality, aiming to solve the of resource scheduling problems in grid computing [14].

(6) Pandey S improved a kind of algorithm based on particle swarm optimization to solve the calculation and transmission system overhead in cloud computing [15].

(7)Hua Xiayu [16] put forward an ant colony optimization scheduling algorithm for cloud computing.

As above, the algorithm is one of the major ways to solve the load balance problems. Some behaviors of animals and plants are simulated and so does the algorithm we used: here we use bees' behavior as the object to illustrate our thought while we use flowers to indicate the VMs we got.

What's more, almost every heuristic algorithm has its own consideration about the load balancing problem exists in cloud computing, for the reason that intelligent algorithm could solve some kinds of NP hard problems properly and the existing load balancing problem lays in the kinds.

3 Our Proposed Method

3.1 Parameter Settings

In order to describe the whole designed systems, the VMs and the cloudlets should be established firstly. So, their parameters based on the actual system should also be set in advance. Parameters of a VM show the characteristics of the VM, such as memory, storage, bandwidth and so forth. A task, which has a pre-assigned instruction length and amount of data, is represented by a cloudlet.

The parameters of a VM can be described as follows:

1) The identification of the VM is defined as *vmid*. It is the unique symbol of the VM that is used to distinguish different VMs in the system.

2) *MIPS* are the millions of instructions per second which is used to show the computing speed. It is a fundamental variable to measure the speed of CPUs.

3) The image size is defined as *size*. It is the size of storage in a VM and is used to set available space.

4) The *ram* describes VM memory. As long as the computer is running, operation data will be transferred to the memory. When the tasks are completed, the results will be sent from the ram.

5) The *BW* is used to describe bandwidth. It represents that how much data can be transmitted in a fixed period, and it also shows capacity of transmitting data in the transport pipeline.

6) The number of *cpu* is pes number.

7) The virtual machine manager's name is *VMm*.

The parameters of cloudlet can be described as follows:

1) The cloudlet is marked as *id*. It is the unique symbol of a cloudlet in the overall system, so that we can use it to distinguish different cloudlets.

2) The length of a cloudlet is recorded as *length*. It refers to the size of a cloudlet which the VM operates.

3) The size of a file is described as *filesize*.

4) The size of an output file is *outputsize*. After tasks are completed, it figures the size of the cloudlets that will be output.

5) The utilization of a cloudlet is *utilizationModel*.

3.2 Mathematical Model

The load balancing problem discussed in this paper is based on the number of cloudlets and the capacity of VMs [17].The capability of VM is shown in the following formulation (1).

$$LOAD = \mathrm{MIPS} * Num\mathrm{CPU} + \mathrm{BW} \tag{1}$$

As we can seen above, the *LOAD* shows the capacity of Vm, the MIPS gives millions of instructions per second, the *NumCPU* illustrates the number of cups, and the *BW* presents the bandwidth.

Generally speaking, capacity of a VM load is based on its own capacity of processing tasks where the *MIPS* and the number of CPU indicate. On the other hand, cloud computing cannot do anything without network, so VM's load is also decided by the bandwidth.

There are no doubts that MIPS decides the computing capability of a virtual machine. And the bandwidth presents the amount that cloud systems are affected by the network.

To meet the needs of load balance, we give the definition of it firstly. It is bases on the variance of the tasks' length sum and is scheduled by every VM used. It is shown in the following formulation (2).

$$P_i = F_i / \sum F_i \tag{2}$$

In the above formulation (2), P_i is the percentage of load of a VM in the system. F_i is the length of scheduled tasks of a VM that describes the load.

At last, we introduce a judgment of balance as shown in the following formulation (3).

$$bla = \sqrt{\sum ((\sum P_i) / vmNum - P_i)^2} / vmNum \tag{3}$$

Here, the *vmNum* is the number of all VMs from the sample or the overall system according to the actual circumstance. And the variance represents average rate about the load of a VM accounting to the load of all VMs. The *bla* will be used to estimate the load condition in the current system. Additionally we also control the system with it. Through making the *bla* in some interval, we can balance the system.

3.3 Tasks Scheduling Algorithm

Tasks scheduling algorithm aims at searching reasonable solutions for a tasking assignment system at one time. In other words, it gives a locally optimal solution. It means that the algorithm cannot have access to the optimal solutions for all problems, whereas it can produce a wide range of optimal approximate solutions for the global optimal solutions [18].

The scheduling algorithm in this paper is presented as follows:

1) **Initialization**: Initialize all parameters of VMs and cloudlets. they are set based on the actual environment which we need in our projects.

2) **Selection**: According to the given rate (n%), n% of all VMs are chosen as a sample to imply the NO.i task. At this step, we will choose the optimal VM in all selected VMs by formulation (2). Because of less VMs, the efficiency can be accepted.

3) **Judging balance:** If the system is under the balance situation, we will distribute the task to the chosen virtual machine. Otherwise, go to the next step.

4) **Iteration**: After choosing a VM, as a result of the randomness of the VM selection in the Step 2, we can use the iteration formulation to search an efficient VM to accomplish the task. The iteration formulation by which we can find a much more optimal solution is shown in the following formulation (4).

$$LOAD_i(t+1) = LOAD_i + \phi(LOAD_i(t) - LOAD_k(k)) \tag{4}$$

In the above formulation, the iterative times is represented by t, and ϕ is a random number between 0 and 1. Through the iteration, a better solution that scheduling the task to VMs will be got.

5) **Judging overload**: After computing "*bla*", we can make the judgement that whether the current system is overloaded or not. According to the result, we decide whether to use the limited formulation to balance the system or not. If the system is overloaded, then go to the Step 6. Otherwise go to the Step 8.

6) **Limited**: If the result in the Step 4 shows that current system is overloaded, we will use the formulation below to renewing the system to schedule the task to another VM, as show in the following formulation (5).

$$LOAD_i = LOAD_i + r*(LOAD_{\max} - LOAD_{\min}) \tag{5}$$

In the above limited formulation, r is a random number between 0 and 1. $LOAD_{\max}$ and $LOAD_{\min}$ represent maximum and minimum the load point respectively.

7) **Judging balance**: If the system is not balanced, then judge the loop ends , based on which make a decision to decide whether to go to Step 3 or not. Otherwise go to the Step 8.

8) **Distributing tasks:** In this step, if the system is under the balance situation, we distribute the task to the chosen virtual machine. However, if the system is still not balanced, we will distribute task to a random chosen virtual machine. Because after using the iteration formulation and limited formulation, if there is still not an appropriate virtual machine to finish the task, we can infer the current system has been not balanced.

9) **Judging end**: By judging whether there is another task, we will go back to Step 2 or stop the process.

4 Experimental Results

To show the efficiency and the accuracy of our proposed load balance method based on the ABC algorithm, the CloudSim is used for testing and giving some experimental results.

In the **first experiment**, 350 cloudlets, 30 VMs and 750 cloudlets, 70 VMs are used in the experiments respectively. The experiment is designed to show the stability and randomness of results of the ABC algorithm. Some results are described in the following figure 2.

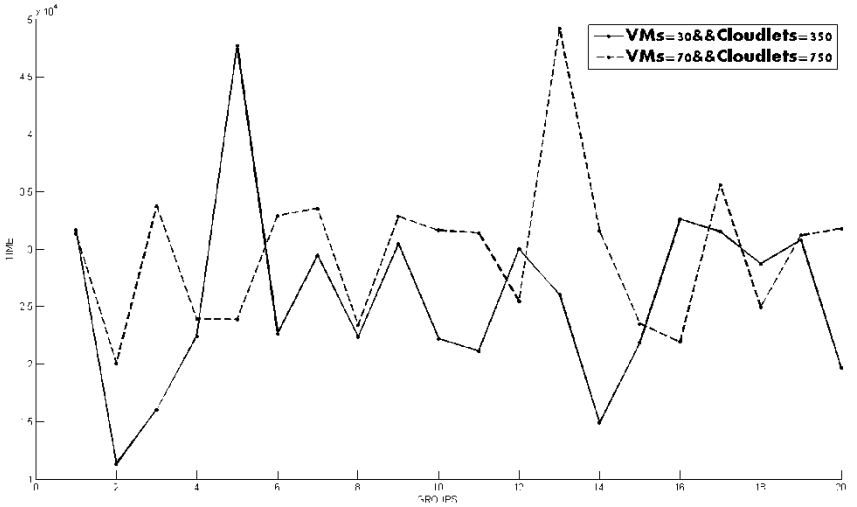


Fig. 2 The results in the first group

The detailed experimental results are shown in the following two tables corresponding to the above two curves in figure 2.

Table 1 Detail data for experiment 1

	20 times testing "finish time" results for 70 VMs and 750 Cloudlets				
Group1(ms)	31349.53	20083.28	33729.2	23945.36	23914.7
Group2(ms)	32920.51	33552.31	23413.74	32866.8	31647.99
Group3(ms)	31413.04	25476.38	49233.06	31578.18	23515.54
Group4(ms)	21933.93	35610.36	24949.06	31209.13	31800.49

	20 times testing "finish time" results for 30 VMs and 350 Cloudlets				
Group5(ms)	31643.31	11309.4	16030.74	22470.5	47741.43
Group6(ms)	22659.38	29474.47	22358.58	30478.46	22208.69
Group7(ms)	21142.57	30029.56	26066.7	14929.83	21855.76
Group8(ms)	32606.21	31522.05	28731.22	30811.19	19660.03

The above graph illustrates that there is a fluctuation between 10000 and 50000. The reason behind is that the algorithm is an intelligent and random algorithm, which does not produce a constant value, in another word, it is a dynamic random algorithm. Particularly , the line for 70 VMs and 750 Cloudlets is above the other. This is because of more system consumption.

The **second experiment** gives the results when the number of cloudlets and the number of VMs change simultaneously. As shown in the following figure 3. And table 2 shows the detailed data in this experiment.

Table 2 Parameters and results in the experiment 2

	Group1	Group2	Group3	Group4	Group5
VMs	10	20	30	40	50
Cloudlets	100	200	300	400	500
Best(ms)	1957.91	4529.31	3808.55	6929.32	6401.9
Worst(ms)	9432.57	24110.44	37572.82	50335.75	63061.89
Average(ms)	3632.91	11347.72	15674.4	24322.16	16134.31

	Group6	Group7	Group8	Group9	Group10
VMs	60	70	80	90	100
Cloudlets	600	700	800	900	1000
Best(ms)	8836.42	18710.71	10001.59	10409.98	11071.48
Worst(ms)	51454.2	33145.77	68567.37	54844.61	24230.93
Average(ms)	19273.42	25819.79	23187.1	19144.77	13457.66

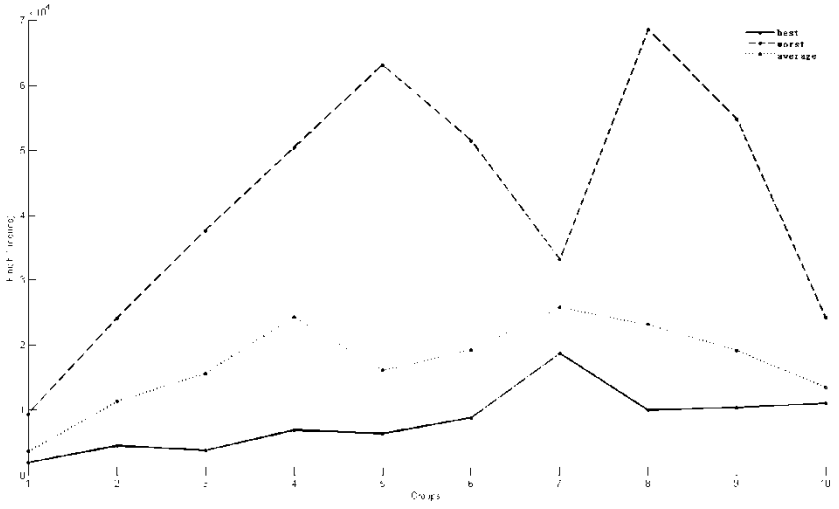


Fig. 3 The results in the second group

Three lines in the above graph represent the best results, the worst results and the average results in ten groups of a large multitude of experiences respectively. In the above figure, when the ratio of cloudlets and VMs is fixed in every group, with the number of cloudlets and VMs increasing simultaneously, all the values fluctuate.

The worst line sharply climbs between the first and fifth group, With a fluctuation from value 20000 to 70000followed. Due to the intelligent algorithm, the worst line shows some random results. However, the best line remains stable during the first and sixth group and then there is a slight fluctuation. Finally, a gentle grow exists in the average line till to the fourth group. Furthermore, the average line decreases slowly after a fluctuation.

5 Conclusion

Based on the ABC algorithm, we propose a load balance method to solve load problems exist in cloud computing. A large multitude of experimental results prove the efficiency of the ABC algorithm.

A load balance method based on the ABC algorithm for cloud computing has been proposed in this paper. This is a reasonable method where the ABC algorithm is used in solving the load balance problems. It can schedule the tasks to reasonable VMs and keep the system under the balanced circumstance. From the sufficient experiments we done, we can draw a conclusion that the algorithm can load the system balance and meet the requirements given from the clients.

Acknowledgements. The authors would like to thank for the support from the project NSFC (National Natural Science Foundation of China) with the Grant number 61202456. We also thank for the Harbin Institute of Technology Innovation Fund project with the Grant number: HIT. NSRIF. 2015087. In addition, we also thank for the support from the Doctor Start Research Project of Shenzhen University (201204).

References

1. Zhu, H., Liu, T., Zhu, D., Li, H.: Robust and simple N-Party entangled authentication cloud storage protocol based on secret sharing scheme. Journal of Information Hiding and Multimedia Signal Processing (JIHMSP) **4**, 110–118 (2013)
2. Chang, B., Tsai, H.-F., Chen, C.-M.: Evaluation of virtual machine performance and virtualized consolidation ratio in cloud computing system. Journal of Information Hiding and Multimedia Signal Processing (JIHMSP) **4**, 192–200 (2013)
3. Zhang, Q., Cheng, L., Boutaba, R.: Cloud computing: state-of-the-art and research challenges. Journal of Internet Services and Applications **1**, 7–18 (2010)
4. Foster, I., Zhao, Y., Raicu, I., et al.: Cloud computing and grid computing 360-degree compared. Grid Computing Environments Workshop, GCE 2008, vol. 1, pp. 1–10 (2008)
5. Vaquero, L.M., Rodero-Merino, L., Caceres, J., et al.: A break in the clouds: towards a cloud definition. ACM SIGCOMM Computer Communication Review **39**, 50–55 (2008)
6. Jadeja, Y., Modi, K.: Cloud computing-concepts, architecture and challenges. In: The International Conference on Computing & Electronics and Electrical Technologies, vol. 1, pp. 877–880. IEEE, Nagercoil (2012)

7. Buyya, R., Ranjan, R., Calheiros, R.N.: Modeling and simulation of scalable cloud computing environments and the cloudSim toolkit: challenges and opportunities. In: International Conference on High Performance Computing & Simulation, HPCS 2009, vol. 1, pp. 1–11. IEEE (2009)
8. Calheiros, R.N., Ranjan, R., Beloglazov, A., et al.: CloudSim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms. Software: Practice and Experience **41**, 23–50 (2011)
9. Das, S., Viswanathan, H., Rittenhouse, G.: Dynamic load balance through coordinated scheduling in packet data systems INFOCOM 2003. In: Twenty-Second Annual Joint Conference of the IEEE Computer and Communications, vol. 1, pp. 786–796. IEEE Societies, IEEE (2003)
10. Braun, T.D., Siegel, H.J., Beck, N., et al.: A comparison of eleven static heuristics for mapping a class of independent tasks onto heterogeneous distributed computing systems. the. Journal of Parallel and Distributed computing **61**, 810–837 (2001)
11. Cañón, J., Alexandrino, P., Bessa, I., et al.: Genetic diversity measures of local European beef cattle breeds for conservation purposes. Genetics Selection Evolution **33**, 311–332 (2001)
12. Jijian, L., Longjun, H., Haijun, W.: Prediction of vibration response of powerhouse structures based on LS-SVM optimized by PSO. Engineering Sciences **12**, 009 (2011)
13. Kazem, A., Rahmani, A.M., Aghdam, H.H.: A modified simulated annealing algorithm for static task scheduling in grid computing. In: International Conference on Computer Science and Information Technology, ICCSIT 2008, vol. 1, pp. 623–627. IEEE (2008)
14. Yulan, J., Zuhua, J., Wenrui, H.: Multi-objective integrated optimization research on preventive maintenance planning and production scheduling for a single machine. International Journal of Advanced Manufacturing Technology **39**, 954–964 (2008)
15. Pandey, S., Wu, L., Guru, S.M., et al.: A particle swarm optimization-based heuristic for scheduling workflow applications in cloud computing environments. In: 2010 24th IEEE International Conference on Advanced Information Networking and Applications (AINA), vol. 1, pp. 400–407. IEEE (2010)
16. Hua, X., Zheng, J., Hu, W.: Ant colony optimization algorithm for computing resource allocation based on cloud computing environment. Journal of East China Normal University (Natural Science) **1**, 127–134 (2010)
17. Babu, L.D., Krishna, P.V.: Honey bee behavior inspired load balance of tasks in cloud computing environments. Applied Soft Computing Journal **13**, 2292–2303 (2013)
18. TSai, P.W., Pan, J.S., Liao, B.Y., et al.: Enhanced artificial bee colony optimization. The International Journal of Innovative Computing, Information and Control **5**, 5081–5092 (2009)

Interference Avoidance Function Research of Spread Spectrum System Using Composite Sequence

Bing Zhao, Zuo Li and Fei Xu

Abstract Facing with some problems of there being more narrow-band interference than Anti interference tolerance of Communication system in communication channel, such as the serious deterioration of the communication system and the decreased communication quality. We put forward that generating a new mW sequence by composite m sequence and Walsh sequence. When we use mW sequence to change the cycle of m sequence and the line order of Walsh sequence, the position of composite sequence's power spectrum in frequency band can be changed too. Using this we can avoid narrow-band interference in communication channel, and improve the reliability and efficiency of communication systems. By the MATLAB simulation of using the mW composite sequence as spread spectrum code, using single frequency interference to simulate narrow-band interference, and contrast on its power spectrum and bit error rate, we can illustrate that mW composite sequence can significantly improve the function of communication system.

Keywords Composite sequence · Power spectrum · Direct sequence spread spectrum · Interference avoidance

1 Introduction

Electronic products have become to normal with the development of economy and the improvement of people's life standard, at the same time, the interference in wireless channel is becoming more and more complex. Time domain processing and transform domain processing are common technologies in interference restraining. Narrow-band interference restraining in time domain is easy to be

B. Zhao(✉) · Z. Li · F. Xu
Electronic Engineering, Heilongjiang University, Harbin, Heilongjiang, China
e-mail: zb0624@163.com

T.T. Zin et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 388,
DOI: 10.1007/978-3-319-23207-2_4

realized, but most of its arithmetic need much longer time to keep stable, and it does not act well in instantaneity and convergence speed, so it can only dispose with stable narrow-band signal [1]. Interference restraining in transform domain detects the position of narrow-band interference restraining in frequency spectrum, then multiplies interference signal's frequency spectrum and cutting function to realize interference restraining [2]. If the interference appears in useful signal bandwidth, although we can remove narrow-band interference, much useful signal energy will be lost so that serious distortion will exist in direct spread spectrum signal.

In this article, we use mW composite sequence which is composited by different ordinal Walsh function sequence and m sequence as spread code, and build BPSK spreading-spectrum communication systems by MATLAB, in the environment of white gaussian noise, by analysing the position of power spectrum of modulated signal spread by different Spread spectrum code and narrow-band interference in frequency spectrum, explain that Walsh function sequence and m sequence dodge well in narrow-band interference, so it can improve the validity and reliability in information transmission.

2 Pseudo Random Sequence

Pseudo random sequence is a kind of certain sequence which has random characteristics. They are generated by shift register, while they have a kind of random characteristics. Because of its characteristics, we can't make a judgment from a sequence's characteristics about whether it is pseudo random sequence or real random sequence. We can just judge rely on the way that how a sequence is generated. Pseudo random sequence has good random characteristics, and it is almost the same as the correlation function of white noise, It can be determined in advance and it is repeatable. These characteristics make pseudo random sequence be wide used especially in some key problems of CDMA.

2.1 The Generation of m Sequences

m sequences is the abbreviation of the longest linear feedback shift register sequence, it is generated by add the feedback on shift register, it is a king of pseudo random sequence and has sharp autocorrelation. Figure 1 shows the structure of N linear feedback shift register is as figure 1 showed, $a_0, a_1, \cdots, a_{n-2}, a_{n-1}$ are the initial state of shift register, $c_0, c_1, \cdots, c_{n-2}, c_{n-1}$ are feedback coefficients of shift register, $d(t)$ is the output m sequence, $c_i = 0$ means there is no feedback, disconnecting the feedback line; $c_i = 1$ means there is feedback, connecting the feedback line [3].

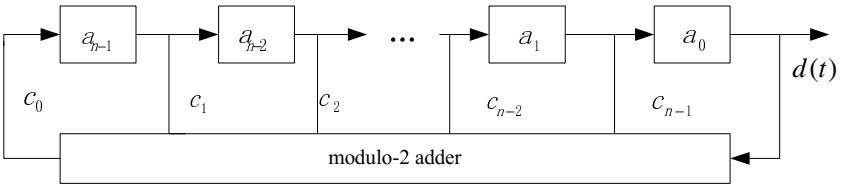


Fig. 1 The structure of N linear feedback shift register

2.2 Walsh Sequences

Walsh sequence is transformed by Hadamard matrix orthogonal square, a kind of Hadamard matrix's Recursive relation whose order $N = 2^n$ is as (1) showed

$$H_n = \begin{bmatrix} H_{n-1} & H_{n-1} \\ H_{n-1} & -H_{n-1} \end{bmatrix} \tag{1}$$

Each line of Hadamard matrix can be seen as a binary sequence, so square H_r, whose order $N = 2^r$ can have *N* sequences, the length of each sequence is *N* too, The concentration of any sequence of two sequences are orthogonal to each other. The cross-correlation function between each sequence which satisfies(2) this sequence is called Walsh sequences [4].

$$R_{ab}(\tau) = \sum_{n=0}^{N-1} a_i b_i = 0 \tag{2}$$

Walsh [5] function sequence matrix is column as rate order, it is called Walsh order. Where I is the Walsh order matrix column rate. When Walsh function sequence matrix is column as rate order, In the positive of don't consider half axis under the condition of spectrum, the center frequency of Walsh sequence (the maximum amplitude of Walsh sequence in frequency spectrum) is determined by the column rate, sequences with same column rate have same spectrum structure. *k* is the line number of Walsh sequences, when *k* is a even number, $i = k/2$; when *k* is an odd, $i = (k+1)/2$. In case of cycle of Walsh sequence is *N*, code width is T_w , Symbol rate is f_w , The center frequency of the sequence is:

$$f_k = i \times \frac{f_w}{N} = \begin{cases} \frac{k+1}{2} \times \frac{f_w}{N}, k = 1,3,5,\cdots \\ \frac{k}{2} \times \frac{f_w}{N}, k = 0,2,4\cdots \end{cases} \tag{3}$$

In addition $Wal(0)$ is DC signal, Line interval sequence between k is 2^p and 2^{p-1} is:

$$\Delta f = 2^p \times \frac{f_w}{N} \qquad (p = 1, 2, \cdots, \log_2 N) \tag{4}$$

Its spectral line of the sequence is $\Delta f = 2 \times \frac{f_w}{N}$

3 Composite Sequences

The composite sequence is a new sequence formed by two or more than two sequences according to some special rules, the sequence which joins in operations called subsequence. In this article, we chose a new mW composite sequence formed by a $Wal(k)$ sequence and a m sequence, the cycle of the $Wal(k)$ sequence is M, the cycle of the m sequence is *N*. In case of they have the same code width, Each element of m sequence is repeated *M* times then forms A sequence, each element of $Wal(k)$ sequence is repeated *N* times then forms B sequence. Then one of the two elements corresponding with the other to get the new sequence by modulo two arithmetic, so it is mW composite sequence [6]. mW composite sequence like this can be to $Wal(k)$ as the unit, according to each symbol in the m sequence values or anti reverse, a number of which are connected in series.

In case of the cycle of the mW sequence is T_{mW}, the cycle of the m sequence is T_m, the code width of Walsh sequence is T_W,

then

$$T_{mW} = M \bullet T_m = M \bullet N \bullet T_W \tag{5}$$

So

$$f_{mW} = f_m / M = f_W / (M \bullet N) \tag{6}$$

put formula (6) in formula (3)

$$f_k = \begin{cases} \dfrac{k+1}{2} \times M \times f_{mW} & k = 1, 3, 5, \cdots \\ \dfrac{k}{2} \times M \times f_{mW} & k = 2, 4, 6, \cdots \end{cases} \tag{7}$$

From formula (7) we can see that The main lobe of the center frequency of mW composite sequence keeps changing with *k* sequence , we can change the position of fixed length composite sequence in the main lobe in the power spectrum by changing the order *k*, So that the main valve can avoid interference and improve the effectiveness and reliability of information transmission.

4 The Composite Sequence of Narrowband Interference Avoidance System

When we use mW composite sequence as spread spectrum code, if interference appears in the main valve position, then the spread spectrum system itself has to play a leading role by its high spreading gain. By changing the line number k we can change the position of mW composite sequence in the power spectrum to avoid the narrow-band interference effect. Based on the assumption of the environment in signal sampler, by Fourier transform we know the frequency of narrow-band interference f_1, carrier frequency of BPSK modulation is f_0, the frequency of mW composite sequence is f_{mW}.

According formula (8)we can figure out the line order k of Walsh sequence.

$$\left\| f_1 - f_0 \right| - f_k \left| \right\rangle \alpha M f_{mW} \qquad \alpha = 1, 2, \cdots, \frac{N}{2} \tag{8}$$

In the formula, α is adjustment coefficient, bigger the α is, better the effect of avoiding. figure 2 shows the block diagram of spread spectrum.

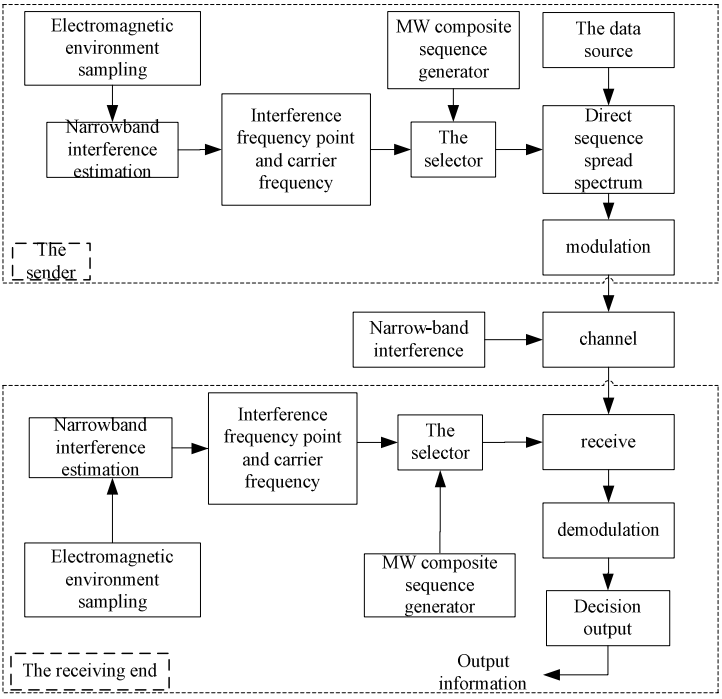


Fig. 2 The block diagram of spread spectrum

4.1 MATLAB Simulation

Narrow-band interference avoidance system of composite sequence use BPSK modulation, carrier frequency is 1000Hz. we chose a new mW composite sequence a $Wal(k)$ sequence and a m sequence as spread spectrum code, the cycle of the m sequence is *M*=7, the cycle of the Walsh sequence is *N*=16. Multiply the spreading code and each symbol of source, then do spread spectrum processing, use single frequency interference simulation with narrow-band interference whose frequency is1000 Hz and SIR is -10dB. When $\alpha = 2$, we can figure out odds whose $k > 3$ or even numbers whose $k > 4$ according to formula (3),(6)and(7),we can chose a mW sequence which is formed by $Wal(5)$ and m sequence and the cycle is 112 to do spread spectrum. When $\alpha = 3$, we can figure out odds whose $k > 5$ or even numbers whose $k > 6$ according to formula (3), (6) and (7), we can chose $Wal(7)$ and m sequence as spread spectrum code to do spread spectrum. Do MATLAB simulation on signal transform in the Gauss white noise channel.

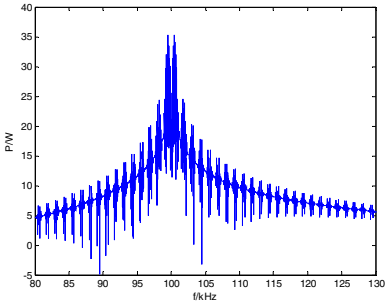


Fig. 3 k=2 power spectrum of modulation signal

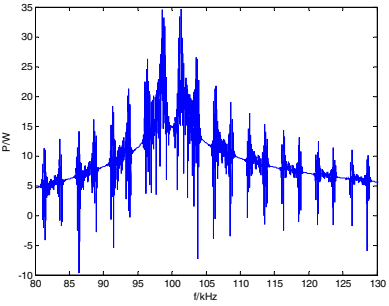


Fig. 4 k=5 power spectrum of modulation signal

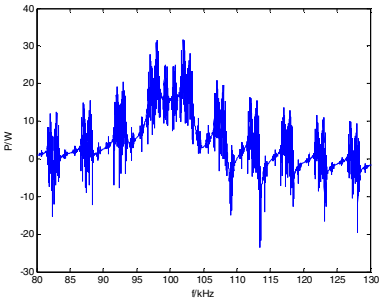


Fig. 5 k=7 power spectrum of modulation signal

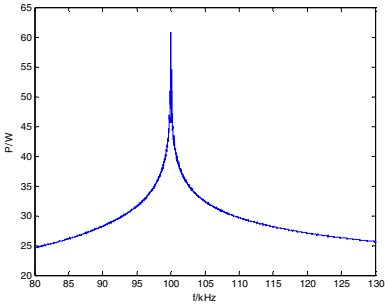


Fig. 6 Narrow-band interference power spectrum

From the figure 3 we can see, using the composite sequence generated by $Wal(2)$ and m sequence as spread spectrum code, its main lobe energy is concentrated in $f_{mW2}=1000$ Hz, from picture 6 we can see that the energy of narrow-band interference concentrated in $f_N=1000$ Hz, composite sequence narrow-band interference avoidance system of $Wal(7)$ has large main lobe interference energy. From figure 4 and figure 5, We can see that the main lobe energy of $Wal(5)$ and $Wal(7)$ composite sequence of narrow-band interference avoidance system distribute on the two sides of $f_{mW2}=1000$ Hz. It avoid the influence of narrow-band interference energy of the main lobe efficiently and reduce the influence of narrow-band interference on the system. And $Wal(7)$ composite sequence of narrow-band interference avoidance system has better function than $Wal(5)$ does. So composite sequence with different order can change the position of Modulation signal in frequency band. Chose right order can avoid narrow-band interference in channel. Though using $Wal(5)$ composite sequence and $Wal(7)$ composite sequence and a m sequence whose cycle is 127 as spread spectrum code, comparing the three sequence when there is no narrow-band interference.

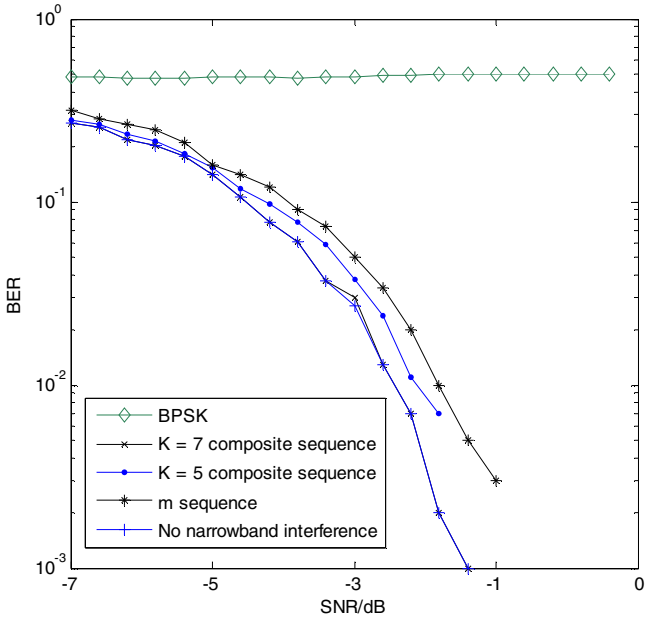


Fig. 7 Curve of bit error rate

We can see from figure 6 that when the SIR is -10dB, BPSK modulation's function can not ensure the normal transmission of information, while the spread

spectrum communication system is still working. Bit error rate of composite sequence when k=7 coincide with BER curve of No DSSS narrow-band interference. It confirms that it can avoid interference effectively. From the comparison between using m sequence as spread spectrum code and the spread spectrum communication system we know that composite sequence has a lower bit error rate than m sequence when they were used as the spread spectrum, it improves the reliability and validity of information transmission.

5 Conclusion

This article composite m sequence with Walsh sequence who has different line order, by analysis on narrow-band interference and Power spectrum of composite sequence with different line order, we know that different line order can change the position of the main lobe spread spectrum signal in power sequence. So that we can avoid the energy of the main lobe of narrowband interference effectively, reduce the impact on the spread spectrum signal of the main lobe and the loss of spread spectrum in transforming. From its BER curve we can see that when composite sequence is used as spread spectrum code, its BER is much lower than conventional way of DS. It can improve the effectiveness and reliability of the communication system effectively, enhance the performance of the communication system.

Acknowledgements. This work is supported by Heilongjiang Provincial Education Department Science and Technology Research Project (NO.12531492). Many thanks to the anonymous reviewers, whose insightful comments made this a better paper.

References

1. Zhan, X., Zhong, S.: Time domain interference cancellation and RAKE energy receiving ultra wideband narrowband interference suppression algorithm based on. J. Science Technology and Engineering **13**, 4182–4185 (2013)
2. Chunhai, Z., Lijun, X., Eryang, Z.: Narrow band interference suppression in transform domain based on adaptive multi threshold algorithm. Jounral of Electronics & Information Technology **28**, 462–465 (2006)
3. Hongxia, Z., Shui, L., Jinyuan, T., Honglie, L.: Simulation of m sequence generation and performance based on Matlab. J. Science & Technology Information **33**, 881–887 (2009)
4. Siwei, L., Huipin, C.: Walhs code spread spectrum communication based on broadband signal. J. Communications Technology **2**, 75–79 (2000)
5. Lili, G., Xin, Z., Jihua, L., Shu, Z.: Walsh (Walsh) analyzing the characteristic of the frequency spectrum of real code. Journal of Harbin Engineering University **5**, 553–555 (2003)
6. Yonghai, W., Lili, G.: MW composite sequence in spread spectrum communication in the application of resisting near-far effect. Journal of Harbin Engineering University **5**, 73–77 (2001)

An Adaptive Kelly Betting Strategy for Finite Repeated Games

Mu-En Wu, Hui-Huang Tsai, Raylin Tso and Chi-Yao Weng

Abstract Kelly criterion is the optimal bidding strategy when considering a series of gambles with the wining probability p and the odds b. One of the arguments is Kelly criterion is optimal in theory rather than in practice. In this paper we show the results of using Kelly criterion in a gamble of bidding T steps. At the end of T steps, there are W times of winning and L times of losing. i.e. $T = W + L$. Consequently, the best strategy for these bidding steps is using the probability W/T instead of using p in Kelly Criterion. However, we do not know the number of W, to put it better the information of p, before placing the bet. We first derive the relation of profits between using p and W/T as the winning probability in the Kelly formula, respectively. Then we use the proportion of winning and bidding numbers before time step t, denoted as p_t, as the winning probability used in the Kelly criterion at time step t. Even we do not know the winning probability of p in a gamble, we can use this method to achieve the profit near the optimal profit when using p in the Kelly betting.

Keywords Kelly criterion · KL-divergence · Winning probability · Odds · Learning theory

M.-E. Wu(✉)
Department of Mathematics, Soochow University, Taipei, Taiwan
e-mail: mnasia1@gmail.com

H.-H. Tsai
Department of Finance, National United University, Miaoli, Taiwan
e-mail: hhtsai@nuu.edu.tw

R. Tso
Department of Computer Science, National Chengchi University, Taipei, Taiwan
e-mail: raylin@cs.nccu.edu.tw

C.-Y. Weng
Department of Computer Science, National Tsinghua University, Hsinchu, Taiwan
e-mail: chiyao.weng@gmail.com

T.T. Zin et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 388,
DOI: 10.1007/978-3-319-23207-2_5

1 Introduction

Kelly criterion, proposed by John Larry Kelly, Jr. [5] at 1956, has been applied in the solution of searching communication optimization and further in those of many aspects, such as the well-known BlackJack [2], Texas hold'em Poker (Hold'em or Holdem in short) and money management of trading financial instruments, including stocks, futures, options and currencies.

Kelly criterion could be regarded as the optimization process of wagering ratio in the long term. Given a set of winning probability and odds, if the gamble could repeat unlimited times, the balance will grows at the fastest speed with the wagering ratio computed by Kelly criterion. However, in the real gamble, we cannot play again and again infinitely, so the interesting question we want to ask here is shown as follows:

Facing the limited times of gambles, is betting by Kelly criterion still our best solution?

Additionally, the condition of Kelly formula applies to is the fixed winning probability and odds, but that of the real gamble or trade is totally different. Ralph Vince [6], [7], [8] provided the coping strategy: the optimal f concept. It optimizes the wagering ratio depending on different odds and then let Kelly formula turn into a special case of this concept. Further, Vince proposed the Leverage Space Model [5] for the staking system similar to Kelly formula in multiple independent gambles, which can optimize the wagering ratio and then to enhance the gambler's profit significantly. Gary. Gottlieb [3] also considered repeated games for Kelly betting. However, he considered an infinite sequence of wagers, which may not practical in real life.

Due to the fact that the situation Kelly formula assumes is a long term of profit maximization process, but the gambles and trades cannot be repeated infinitely, this study tries to investigate the gap between the theory and the reality. We adopt the concept of online learning to computes the relation between the profit from Kelly formula and that from hindsight after T games each. The hindsight here means that after T times of bidding steps, if the wining times is W, then we can replace the original winning probability of p in Kelly formula with W/T. By our derivation, the relation of profits between wagering by Kelly formula and by hindsight can be described by Kullback-Leibler divergence (short for KL divergence in short) [7]. In other words, if the winning probability used in Kelly formula matches that computed from the winning times within T round of bidding steps, then the profits from these two staking strategies will be almost the same as T goes to infinity. The more different between the winning times within T round of bidding steps and the given probability p, the more between profits of these two strategies.

The second achievement in this study shows that under the condition with unknown winning probability, we adopt the already known number of wining rounds divided by the number of bidding numbers as the winning probability needed in Kelly formula to wager, and then the following profit profile of this strategy compared to that of adopting real probability, i.e. hindsight, can be described by KL divergence.

The remainder of this paper is shown in the following: In Section 2, we give preliminaries, include Kelly criterion and KL divergences. In Section 3, the relation between traditional Kelly criterion and hindsight is described. Section 4 analyze the variant of Kelly criterion in the case of unknown winning probability. Finally, we conclude in Section 5.

2 Preliminaries

2.1 Kelly Criterion

Considering a gamble with the winning probability p and odds b. Let a gambler with the initial capital being A_0, the t-th step capital being A_t, and the wager ratio of f, where $0 < f < 100\%$, the Kelly formula can be derived as follows:

If the gambler wins the $(t-1)$-th round, then $\mathrm{A}_t = \mathrm{A}_{t-1}(1+bf)$.

If the gambler loses the $(t-1)$-th round, then $\mathrm{A}_t = \mathrm{A}_{t-1}(1-f)$.

Since the gambler plays T rounds and has the winning probability of p, we can expect that he/she will win $T \times p$ rounds and lose $T \times (1-p)$ rounds. Then in theory, the value of A_T should be:

$$\mathrm{A}_T = \mathrm{A}_0(1+bf)^{Tp}(1-f)^{T(1-p)}$$

By the above equation, we can optimize A_T to find the solution of f. Differentiate the above equation, we can find the optimal f value as following:

$$f = \frac{p(1+b)-1}{b}$$

However, the numbers of winning and losing should depend on binomial distribution during the process of the real T rounds. For example, among 100 time of gambles with the winning probability 50%, it is not just composed of 50 time of wins and 50 time of loses. According to the binomial theorem, there will be the probability of $C_k^{100}50\%^k \times 50\%^{100-k}$ to win k times and to lose $100-k$ times. There is a little difference between theory and practice.

Hence, the issue we want to investigate in this paper is described in the following: Within a sequence of T-time gambles, there are W times of wins and L times of losses, what is the relation between the wagering profit from Kelly formula with the winning probability p and that with the hindsight probability W/T?

2.2 KL Divergence

In this paper, we use KL divergence [4] to describe the difference of Kelly profits between theoretical expectation and practice. KL divergence measures the distance between its two input distributions [1]. It is a non-symmetric measure.

Assuming that P and Q are two discrete distributions, we will have $\mathrm{KL}(P||Q) = \sum_{i=1}^{n} p_i \ln \frac{p_i}{q_i}$. Note that $\mathrm{KL}(P||Q) = 0$ as $P = Q$.

3 On the Relation of Kelly Criterion on Theory and Practice

First, let us define the so-called Hindsight here. During the T-time steps in a gamble, even knowing the winning probability p, we can only compute the probability of winning W times within T time steps of the gamble, where W could be 1, 2, …, T, etc. In fact W is a random variable and will be found after the end of T-time real gambles. No one can know it in advance. Therefore, should we know it in advance, we can substitute W/T into the probability in Kelly formula, so as to use this wager ratio:

$$f^* = \frac{\frac{W}{T}(1+b)-1}{b}$$

Such kind of result will be better than that of the traditional Kelly formula's wager ratio:

$$f = \frac{p(1+b)-1}{b}$$

This is why we call the use of f^* in Kelly formula for wagering propositionally as "Hindsight". Next, supposing that using f and f^* as the wager ratios to play T times and the following results are $\mathrm{E}_T(f)$ and $\mathrm{E}_T(f^*)$ respectively, we describe the relationship between $\mathrm{E}_T(f)$ and $\mathrm{E}_T(f^*)$ in the next theorem:

Theorem 1. Consider a gamble with the winning rate of p and the odds of b, if we wager f ratio of money by the Kelly rule, i.e., $f = (p(1+b)-1)/b$, and continue T time steps. We denote the expected payoff $\mathrm{E}_T(f)$. Further, if there are W times of wining and L times of losing (i.e. $T = W + L$) during the T time steps, the best hindsight is to wager f^* ratio of his money each time steps and get the payoff $\mathrm{E}_T(f^*)$, where f^* equals $(p(1+b)-1)$. Then we have the following equation:

$$\frac{1}{T} \times \ln \frac{\mathrm{E}_T(f^*)}{\mathrm{E}_T(f)} = \mathrm{KL}\left(\frac{W}{T}||p\right)$$

Before the proof of Theorem 1, we observe that the KL divergence of W/T and p are positive related to the ration of $\mathrm{E}_T(f^*)$ and $\mathrm{E}_T(f)$. As W/T approaches the winning rate p sufficiently, $\mathrm{E}_T(f)$ will converge to $\mathrm{E}_T(f^*)$ significantly. This means that although the winning rate p in the Kelly rule works well, the best hindsight is to adopt the alternative winning rate, that is W/T. One thing deserves to mention is that the difference of profit/loss (P/L) between these two betting methods will depends on the KL Divergence between W/T and p.

The Proof of Theorem 1

Within a T-time gambles, we denote the number of winnings as W and the one of losings as L. It is trivial to say that $T = \mathrm{W} + L$. On one hand, according to Kelly criterion, we may set the optimal $f = \frac{p(1+b)-1}{b}$, and then wager by it. The profit/loss, $\mathrm{E}_T(f)$, after T-time of gambles can be computed as follows:

$$\mathrm{E}_T(f) = (1+bf)^W \times (1-f)^L$$

$$= \left(1 + b\frac{P(1+b)-1}{b}\right)^W \times \left(1 - \frac{P(1+b)-1}{b}\right)^L$$

$$= \left(p(1+b)\right)^W \times \left(\frac{1+b}{b}(1-p)\right)^L$$

$$= \left(p(1+b)\right)^W \times \left(\frac{1+b}{b}(1-p)\right)^L$$

On the other hand, because the T-time of gambles are composed of W-time of winnings and L-time of losings, the best staking strategy is to set the winning probability as W/T. Before the end of the T rounds of a gamble, no one will be able to make sure absolutely about the number of W. Therefore this is why we call the staking strategy, which uses the W/T as the winning probability, as a hindsight. We want to compute the profit and loss from the hindsight and the traditional $f = \frac{p(1+b)-1}{b}$ throughout the gambles each. We hope to prove that the difference between the profits of wagering with Kelly formula and with hindsight will be not significant. The derivation we did is as follows:

Replacing p with W/T in Kelly formula, we have $f^* = \frac{\frac{W}{T}(1+b)-1}{b}$. Hence, using f^* to play T times will have the profit and loss like:

$$\mathrm{E}_T(f^*) = (1+bf^*)^W \times (1-f^*)^L$$

$$= \left(1 + b\frac{\frac{W}{T}(1+b)-1}{b}\right)^W \times \left(1 - \frac{\frac{W}{T}(1+b)-1}{b}\right)^L$$

$$= \left(\frac{W}{T}(1+b)\right)^W \times \left(\frac{1+b}{b} \times (1-\frac{W}{T})\right)^L$$

Consequence, the ratio of the profits when bidding with f and f^* is

$$\frac{\mathrm{E}_T(f^*)}{\mathrm{E}_T(f)} = \frac{\left(\frac{W}{T}(1+b)\right)^W \times \left(\frac{1+b}{b} \times \left(1-\frac{W}{T}\right)\right)^L}{\left(p(1+b)\right)^W \times \left(\frac{1+b}{b} \times (1-p)\right)^L} = \frac{\frac{W}{T}^W \times \left(1-\frac{W}{T}\right)^L}{(p)^W \times (1-p)^L}$$

We take the loge function and divide by T to get

$$\frac{1}{T} \times \ln\frac{\mathrm{E}_T(f^*)}{\mathrm{E}_T(f)} = \frac{W}{T}\ln\frac{W/T^W}{p^W} + \left(1-\frac{W}{T}\right)\ln\frac{(1-W/T)^L}{(1-p)^L} = \mathrm{KL}\left(\frac{W}{T}||p\right),$$

which proves the Theorem 1.

The aforementioned theorem proves that given a fixed winning probability, even after T round of simulations we cannot make sure of the numbers of

winnings and losings in reality. But by the Law of Larger Number, W/T will approach p. Therefore, as the number of playing is enough, we can see that the difference between profits from p and from the hindsight will not significant. However, when the gap between W/T and p is large enough, the difference between them will be significant afterwards. The ratio of their profit is just the same as the KL divergence of W/T and p.

4 Learning from Probability Before Time Step t

We extend the results of Section 3 to the general case. Consider a gamble with the unknown win rate and the odds of b. If we adopt the happened winning percentage as the winning probability and b as the odds, then we want to study the relation between profits of this new method and that of the hindsight. The process we deduce is as follows.

Supposing the winning times before every time step t being W_{t-1}, we can assign

$$p_t = \frac{W_{t-1}}{t-1}$$

We use p_t as the winning probability at the time step t. Consequently, we take p_t into Kelly formula and then calculate the bidding fraction f_t at time step t. That is,

$$f_t = \frac{p_t(1+b)-1}{b} = \frac{\frac{W_{t-1}}{t-1}(1+b)-1}{b}$$

Assume that the win/loss profile of this T rounds is r1, r2 , r3,…rT, where ri={win, lose}. Therefore, the profit after T rounds is

$$\text{Profit}_T = (1+B_1f_1)\times(1+B_2f_2)\times\ldots\times(1+B_{T-1}f_{T-1})\times(1+B_Tf_T),$$

where $B_i = b$ if r_i is win, and $B_i = -1$ if r_i is lose.

Without loss of generality, we rearrange the sequence of r1, r2 , r3,…rT. Those winning rounds are located at the front and those losing ones are moved to behind them in the same turn in each group. For example, if the sequence of win-lose is

$$(w_1, l_1, w_2, w_3, w_4, l_2, \ldots l_{L-2}, w_W, w_{L-1}, l_L)$$

Rearrange the order of this sequence to

$$(w_1, w_2, w_3, \ldots, w_{W-1}, w_W, l_1, l_2, \ldots, l_{L-2}, l_{L-3})$$

We have

$$\text{Profit}_T = \left(1+bf_{w_1}\right)\times\left(1+bf_{w_2}\right)\times\ldots\times\left(1+bf_{w_W}\right)\times\times\left(1-f_{l_1)}\right)\left(1-f_{l_2}\right) \times\ldots\times\left(1-f_{l_L}\right)$$

Hence,

$$\text{Profit}_T = \left(\prod_{i=1}^{W}\left(1+bf_{w_i}\right)\right)\times\left(\prod_{k=1}^{L}\left(1-f_{l_k}\right)\right)$$

$$= \left(\prod_{i=1}^{W}\left(1 + b\frac{p_{W_i}(1+b)-1}{b}\right)\right) \times \left(\prod_{k=1}^{L}\left(1 - \frac{p_{l_k}(1+b)-1}{b}\right)\right)$$

$$= \left(\prod_{i=1}^{W}\left(p_{W_i}(1+b)\right)\right) \times \left(\prod_{k=1}^{L}\left(\frac{1+b}{b}\right)\left(1-p_{l_k}\right)\right)$$

$$= \frac{(1+b)^T}{b^L}\left(\prod_{i=1}^{W} p_{W_i}\right) \times \left(\prod_{k=1}^{L}\left(1-p_{l_k}\right)\right)$$

Thus,

$$\frac{\mathrm{E}_T(f^*)}{\mathrm{Profit}_T} = \frac{\left(\frac{W}{T}(1+b)\right)^W \times \left(\frac{1+b}{b}\left(1-\frac{W}{T}\right)\right)^L}{\left(\prod_{i=1}^{W}\left(1+bf_{w_i}\right)\right) \times \left(\prod_{k=1}^{L}\left(1-f_{l_k}\right)\right)}$$

$$= \frac{\frac{(1+b)^T}{b^L}\left(\frac{W}{T}\right)^W \times \left(1-\frac{W}{T}\right)^L}{\frac{(1+b)^T}{b^L}\left(\prod_{i=1}^{W} p_{w_i}\right) \times \left(\prod_{k=1}^{L}\left(1-p_{l_k}\right)\right)} = \frac{\left(\frac{W}{T}\right)^W \times \left(1-\frac{W}{T}\right)^L}{\left(\prod_{i=1}^{W} p_{w_i}\right) \times \left(\prod_{k=1}^{L}\left(1-p_{l_k}\right)\right)}$$

Lemma 2. There exist a value p_{WL} satisfying

$$p_{WL} = \left(\prod_{i=1}^{W} p_{w_i}\right) \times \left(\prod_{k=1}^{T-W}\left(1-p_{l_k}\right)\right) = {p_{WL}}^W \times (1-p_{WL})^{T-W}.$$

Proof of Lemma 2. The Lemma is trivial since p_{WL} is just the solution of the polynomial with single variable.

We call p_{WL} the geometric mean of probability. After taking the $\log_e$ and dividing by T, we have

$$\frac{1}{T}\ln\frac{\mathrm{E}_T(f^*)}{\mathrm{Profit_T}} = \frac{1}{T}\ln\frac{\left(\frac{W}{T}\right)^W \times \left(\left(1-\frac{W}{T}\right)\right)^L}{(p_{WL})^W \times (1-p_{WL})^L}$$

$$= \frac{W}{T}\ln\frac{W/T}{p_{WL}} + \left(1-\frac{W}{T}\right)\ln\frac{1-W/T}{1-p_{WL}} = \mathrm{KL}\left(\frac{W}{T}||p_{WL}\right)$$

The above-mentioned result shows that wagering according to the sequential winning percentage, the final profit will depend on the KL divergence of geometric mean of probability and the real winning percentage.

One thing deserves mention is that as the number of T increases, p_{WL} will approach p. In other words, as the rounds of gambles increase, the happened winning percentage will present the unknown winning probability of this gamble. However, if we wager following the ratio computed from the winning percentage happened which used as the winning probability in Kelly formula, so long as the time we play is enough, i.e. T is large enough, and its profit will tend towards that of hindsight.

5 Conclusion and the Future Work

This study first derives that when using Kelly formula, the relation between the real profit and loss and the expected one can be described by KL divergence, and then deduces the same result of KL divergence for describing the close relation between the profits from hindsight and from our adaptive winning probability, which is the winning percentage of the past events under the condition of unknown winning probability in real world in advance, combining the given odds to wager by Kelly formula in the long term. Applying this kind of technique, we hope to use the happened events under the situation with unknown winning probability and odds as the learning sample to predict the possible probability and odds needed for wagering by Kelly formula. Therefore, the study in this paper can be used in stock market prediction, money management and the development of trading strategy.

Acknowledgments. The work of Mu-En Wu has been supported by the Ministry of Science and Technology, Taiwan, R.O.C., under Grant MOST 103-2218-E-031 -001. The work of Raylin Tso has been supported by the Ministry of Science and Technology, Taiwan, R.O.C., under Grant MOST 103-2221-E-004-009.

References

1. Chou, J.-H., Lu, C.-J., Wu, M.-E.: Making profit in a prediction market. In: Gudmundsson, J., Mestre, J., Viglas, T. (eds.) COCOON 2012. LNCS, vol. 7434, pp. 556–567. Springer, Heidelberg (2012)
2. Thorp, E.O.: The kelly criterion in blackjack, sports betting, and the stock market. In: Zenios, S.A., Ziemba, W. (eds.) Handbook of Asset and Liability Management, vol. 1 (2006)
3. Gottlieb, G.: An optimal betting strategy for repeated games. Journal of Applied Probability, 787–795 (1985)
4. http://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence
5. Kelly, J.L.: A new interpretation of information rate. Bell System Technical Journal **35**(4), 917–926
6. Vince, R.: Portfolio Management Formulas. John Willey & Sons, New York (1990)
7. Vince, R.: The mathematics of money management: risk analysis techniques for traders, vol. 18. John Wiley & Sons (1992)
8. Vince, R., The new money management: a framework for asset allocation, vol. 47. John Wiley & Sons (1995)
9. Vince, R.: The leverage space trading model: reconciling portfolio management strategies and economic theory, vol. 425. John Wiley and Sons (2009)

A Sanitization Approach of Privacy Preserving Utility Mining

Jerry Chun-Wei Lin, Tsu-Yang Wu, Philippe Fournier-Viger, Guo Lin, Tzung-Pei Hong and Jeng-Shyang Pan

Abstract High-Utility Itemset Mining (HUIM) considers both quantity and profit factors to measure whether an item or itemset is a profitable product. With the rapid growth of security considerations, privacy-preserving utility mining (PPUM) has become a critical issue in HUIM. In this paper, an efficient algorithm is proposed to minimize side effects in the sanitization process for hiding sensitive high utility itemsets. Three similarity measurements are also designed as the new standard used in PPUM. Experiments are also conducted to show the performance of the designed algorithm in terms of general side effects in PPDM and the new defined measurements in PPUM.

J.C.-W. Lin(✉) · T.-Y. Wu · G. Lin
Innovative Information Industry Research Center(IIIRC), School of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, China
e-mail: jerrylin@ieee.org, {wutsuyang,linguo.hit}@gmail.com

P. Fournier-Viger
Department of Computer Science, University of Moncton, Moncton, Canada
e-mail: philippe.fournier-viger@umoncton.ca

T.-P. Hong
Department of Computer Science and Information Engineering,
National University of Kaohsiung, Kaohsiung 811, Taiwan, R.O.C
e-mail: tphong@nuk.edu.tw

T.-P. Hong
Department of Computer Science and Engineering, National Sun Yat-sen University,
Kaohsiung 804, Taiwan, R.O.C.

J.-S. Pan
College of Information Science and Engineering, Fujian University of Technology,
Fuzhou 350118, China
e-mail: jengshyangpan@gmail.com

T.T. Zin et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 388,
DOI: 10.1007/978-3-319-23207-2_6

1 Introduction

Due to quick proliferation of electronic data from government, corporations and organizations, the discovered knowledge may, however, implicitly contain confidential, private or secure information and lead to privacy threats if they are published or misused [2, 4, 23]. Privacy-Preserving Data Mining (PPDM) was thus proposed to hide the sensitive information by perturbing the original database and producing a sanitized one [3, 7, 10]. As the similar considerations of PPDM, Privacy-Preserving Utility Mining (PPUM) has also become an important topic in recent years. A sensitive high utility itemset indicates that an itemset is a high utility itemset but is concerned as a confidential or secure information and required to be hidden before it is published or shared. Yeh et al. [26] first proposed Hiding High Utility Itemset First (HHUIF) algorithm and Maximum Sensitive Itemsets Conflict First (MSICF) algorithm to hide the sensitive high utility itemsets. Lin et al. presented a GA-based algorithm for hiding sensitive high utility itemsets through transaction insertion [20]. Yun et al. proposed the Fast Perturbation algorithm Using a Tree structure and Tables (FPUTT) algorithm [27] to speed the sanitization process with an aided tree structure and the associated index table. The above approaches use the similar criteria of PPDM to measure the performance of the developed algorithms in sanitization process, which is insufficient for PPUM.

In this paper, an algorithm is proposed to efficiently hide the sensitive high utility itemsets with minimal side effects. Since the previous criteria of PPDM [5] is not suitable to evaluate the performance of the developed algorithms in PPUM, three similarity measurements are also designed in this paper to show the effectiveness and efficiency of the algorithms developed for PPUM. Experiments are then conducted to show that the proposed algorithm has better results compared to the state-of-the-art HHUIF and MSICF algorithms.

2 Related Work

Data mining can be concerned as a powerful way to reveal the implicit relationships among the itemsets from a very large database [1, 6, 12]. In some applications, the private or confidential information is required to be hidden before it is published in public or shared with collaborators. Privacy-Preserving Data Mining (PPDM) has thus become a critical issue in recent years. The purpose of PPDM is to hide sensitive itemsets with minimal side effects. The relationship of itemsets before and after PPDM procedure can be seen in Fig. 1, where F represents the large itemsets of D; S represents the sensitive itemsets defined by users that are large; and F' is the large itemsets after some sanitization process. From Fig. 1, the α (hiding failure) is the set of sensitive itemsets which were failed to be hidden after data sanitization procedure; the β (missing cost) is the set of non-sensitive frequent itemsets in the original database, but could not be mined out from the sanitized database; and the γ (artificial cost) is the set of frequent itemsets appearing in the sanitized database but not frequent ones in the original database.

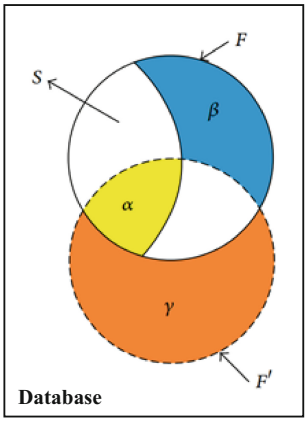


Fig. 1 Three side effects of PPDM.

Verykios et al. proposed three strategies with five designed approaches to decrease support or confidence for hiding sensitive association rules [22]. Sun et al. proposed a border-based method to sanitize the database for hiding sensitive frequent itemsets [21]. Li et al. proposed a kd-tree approach to recursively perturb the partitioned sub-databases for PPDM [14]. Li et al. proposed MICF [15] algorithm to select the victim item with the maximum conflict degree in a sensitive transaction. Hong et al. proposed a SIF-IDF algorithm to hide sensitive itemsets based on the TF-IDF mechanism [13]. Lin et al. proposed a GA-based approach to hide the sensitive frequent itemsets with transaction deletion [19].

High Utility Itemset Mining (HUIM) is an extension of frequent itemset mining which considers both the quantity and profit factors to find profitable high-utility itemsets (HUIs) from a transactional database [24, 25]. Several algorithms have been extensively studied to efficiently mine the HUIs [17, 18, 24]. Since HUIM has become an emerging topic in variant applications, Privacy-Preserving Utility Mining (PPUM) has also become a critical issue in recent years. Yeh et al. first designed the HHUIF and MSICF algorithms to hide SHUIs [26] in PPUM. Lin et al. first developed a GA-based method to hide the sensitive high utility itemsets [20] by inserting the dummy transactions. The above approaches only concern the similar three side-effects of PPDM to evaluate the efficiency of the developed algorithms, which is not sufficient in PPUM. In this paper, we develop a sanitization algorithm and three utility-similarity measurements as the new standard to evaluate the effectiveness and efficiency for hiding the sensitive high utility itemsets in PPUM.

3 Preliminaries and Problem Statement

Given a finite set of items $I = \{i_1, i_2, \ldots, i_k\}$, each item i_j has its own positive unit profit as $p(i_j)$. An itemset X is a set of k distinct items in which $X \in I$. A quantitative

database is denoted as $D = \{T_1, T_2, \ldots, T_n\}$. Each item i_j in transaction T_q is associated with its quantitative value as $q(i_j, T_q)$. A quantitative example and its profit table are used as the running examples and respectively given in Tables 1 and 2.

Table 1 A transactional database.

TID	Transaction (item, quantity)
1	$(A, 2), (B, 1), (E, 3)$
2	$(C, 1), (D, 6)$
3	$(B, 1), (C, 2), (E, 1), (F, 1)$
4	$(A, 3), (B, 4), (C, 2), (D, 2), (E, 5)$
5	$(B, 3), (C, 5)$
6	$(A, 2), (E, 7), (F, 3)$

Table 2 A profit table.

Item	Profit
A	5
B	3
C	2
D	1
E	6
F	10

Definition 1. *The utility of an item i_j in a transaction T_q is denoted as $u(i_j, T_q)$, which can be defined as:*

$$u(i_j, T_q) = p(i_j) \times q(i_j, T_q). \tag{1}$$

Definition 2. *Let $X = \{i_1, i_2, \ldots, i_k\}$ be an itemset. The utility of an itemset X in a transaction T_q is denoted as $u(X, T_q)$, which can be defined as:*

$$u(X, T_q) = \sum_{i_j \in X \wedge X \in T_q} u(i_j, T_q). \tag{2}$$

Definition 3. *The utility of an itemset X in D is denoted as $u(X)$, which can be defined as:*

$$u(X) = \sum_{X \in T_q \wedge T_q \subseteq D} u(X, T_q). \tag{3}$$

Definition 4. *An item/set is defined as a High-Utility Itemset (HUI) if its utility is no less than the user-defined minimum utility threshold δ as:*

$$HUI \leftarrow \{X | u(X) \geq \delta\}. \tag{4}$$

Problem Statement: Given a set of the sensitive high utility itemsets to be hidden as $\{s_1, s_2, \ldots, s_m\}$, in which each $s_d \in$HUI. The problem statement of Privacy-Preserving Utility Mining (PPUM) is to completely hide the sensitive high utility

itemsets until their utilities are less than the pre-defined minimum utility threshold δ. For PPUM, it is insufficient to reveal the traditional three side-effects used in PPDM. The utility factor should also be involved as the evaluation criteria to show the efficiency and effectiveness for the algorithms developed in PPUM .

4 Proposed Sanitization Algorithm

In this paper, an efficient algorithm is proposed to hide the pre-defined sensitive high utility itemsets with minimum utility consideration. The designed algorithm is described in Algorithm 1.

Algorithm 1. Proposed algorithm

Input: D, the quantitative database; *ptable*, the profit table; *SHUIs*, the set of sensitive high-utility itemsets to be hidden; δ, minimum utility threshold.
Output: A sanitized database D'.

build an index *iTable* ;
for *each* $s_d \in SHUIs$ **do**
  find numbers of i_j as $f(i_j), i_j \in s_d$;
  sort s_d in descending order of $f(i_j)$;
for *each* $s_d \in SHTUIs$ **do**
  $du(s_d) := u(s_d) - \delta$;
  if $du(s_d) < 0$ **then**
    continue;
  else
    project $D_{s_d} \leftarrow (D, s_d, iTable)$;
    sort D_{s_d} in descending order of $u(s_d, T_d)$;
    for $i := 1$ *to* $\left|D_{s_d}\right|$ **do**
      $T^{vic} \leftarrow \{T_d | d = i, T_d \in D_{s_d}\}$;
      $i^{vic} \leftarrow min\{u(i_k, T^{vic}), 1 \le k \le \left|s_d\right|, i_j \in s_d\}$;
      if $u(i^{vic}) < du(s_d)$ **then**
        delete i^{vic} from T^{vic};
        $du(s_d) := du(s_d) - u(i^{vic}, T^{vic})$;
        update $SHUIs, iTable, D_{s_d}$;
        if $du(s_d) \le 0$ **then**
          break;
      else
        $dqvalue := \left\lfloor \frac{u(i^{vic})}{p(i^{vic})} \right\rfloor$;
        $q(i^{vic}, T^{vic}) := q(i^{vic}, T^{vic}) - dqvalue$;
        $tu(T^{vic}) := tu(T^{vic}) - u(i^{vic})$;
        delete s_d from $SHUIs$;
        update $iTable, D_{s_d}$;
        break;

5 Experimental Evaluation

From the conducted experiments, a real foodmart dataset [11] and a synthetic T25I10D10K [9] dataset are used to evaluate the performance of the proposed algorithm compared to those of the state-of-the-art HHUIF and MSICF algorithms [26]. Since the algorithms used in PPUM have different considerations compared to the algorithms used in PPDM, three similarity measures namely Database structure similarity (DSS), Database Utility Similarity (DUS), and Itemsets Utility Similarity (IUS) are designed as the novel criteria to evaluate algorithms developed in PPUM.

5.1 Runtime

The execution time of three algorithms under varied dataset sizes in two datasets are compared and shown in Fig. 2.

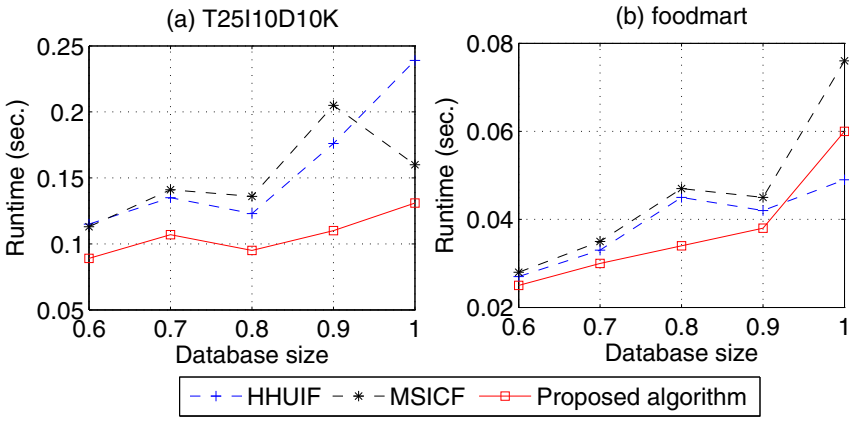


Fig. 2 Runtime w.r.t varied dataset sizes.

From Fig. 2, it can be seen that when the dataset size is increased, more computations are required to hide the sensitive high utility itemsets with the increasing of their utilities. The reason is that when the dataset size is larger, the utilities of the sensitive high utility itemsets are also increased. Thus, it is necessary to require more times to find the appropriate transactions to delete the SHUIs or decrease the utilities of them. The conducted experiments in Fig. 2 also indicates that the proposed algorithm always has better results compared to the state-of-the-art HHUIF and MSICF algorithms.

5.2 *Database Structure Similarity*

A Database Structure Similarity (DSS) measurement is first designed in this paper to evaluate the structure similarity before and after sanitization process, which indicates the number of modified transactions in the sanitization process. The designed evaluation criteria of DSS is given below.

Definition 5. *Let D and D' are respectively the original database and the sanitized database. The pattern of a transaction tp_k is represented as $\{i_1, i_2, ..., i_m\}$, in which m is the number of items in the database D and i_j is represented as 1 if it appears in the transaction; otherwise, it is represented as 0. The DSS criteria is thus defined as:*

$$DSS = \sqrt{\sum_{k=1}^{|tp_k^D \cup tp_k^{D'}|} (freq(tp_k^D) - freq(tp_k^{D'}))^2}, \tag{5}$$

where $freq(tp_k^D)$ is the frequency of the represented pattern in the original database D, and $freq(tp_k^{D'})$ is the frequency of the represented pattern in the perturbed database D'. The DSS results of four algorithms under varied dataset sizes and sensitive percentages in four datasets are compared and respectively shown in Fig. 3.

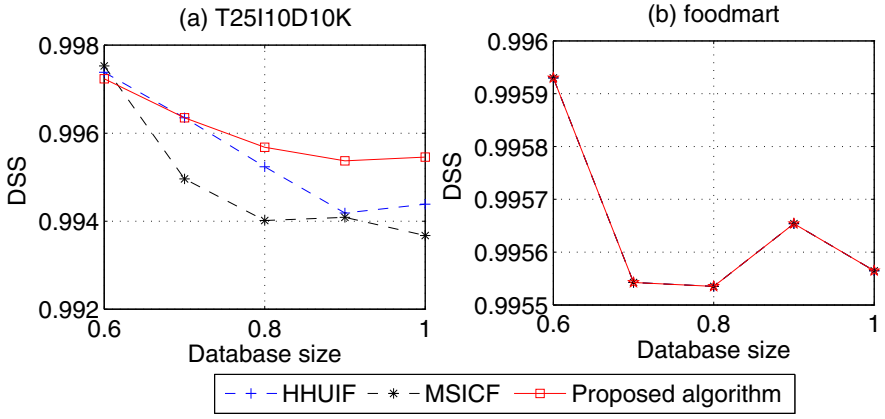


Fig. 3 Database structure similarity (DSS) w.r.t. varied dataset sizes.

In Fig. 3(b), all algorithms have the same results in DSS criteria. From Fig. 3(a), it can be seen that the proposed algorithm generally has better performance compared to the HHUIF and MSICF algorithms.

5.3 Database Utility Similarity

The Database Utility Similarity (DUS) is used to measure the degree of utility loss in the entire dataset. This criterion can be used to reveal the actually reduced utility (utility loss) in the sanitization process, which is more suitable in PPUM. The evaluation equation is then shown below.

Definition 6. *Let D, D' are the original database and the sanitized database respectively. The loss utility between the original database and the sanitized database is denoted as Database Utility Similarity (DUS), which can be defined as:*

$$DUS = \frac{\sum_{T_q \in D'} tu(T_q)}{\sum_{T_q \in D} tu(T_q)}. \tag{6}$$

The DUS results of three algorithms under varied dataset sizes in two datasets are compared and shown in Fig. 4. From Fig. 4, it can be seen that the proposed algorithm has the best results compared to the other two algorithms in two datasets. The reason is that the minimum utility mechanism is adopted in the developed algorithm, fewer utility loss may require to hide the sensitive high itemsets.

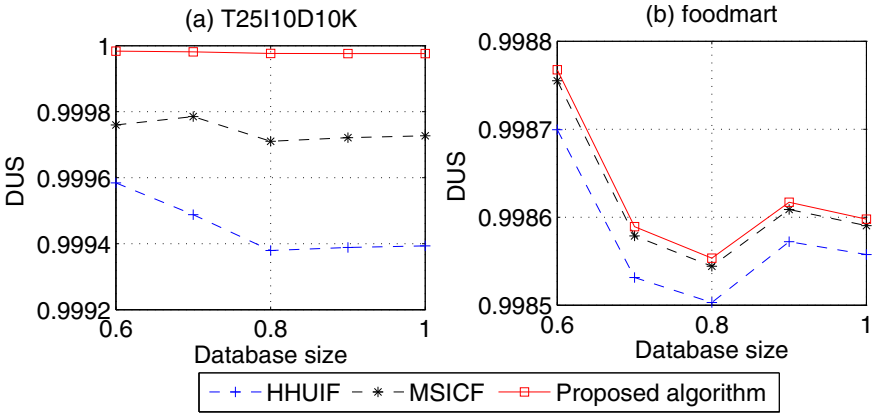


Fig. 4 Database utility similarity (DUS) w.r.t. varied dataset sizes.

5.4 Itemsets Utility Similarity

In addition to the designed DUS criteria to evaluate the performance of the developed algorithms in PPUM, an Itemset Utility Similarity (IUS) is also designed as another measurement to show the loss utilities of the discovered HUIs before and after sanitization process in PPUM. The evaluation equation is then shown below.

Definition 7. *Let $HUIs^D$ and $HUIs^{D'}$ are the discovered high-utility itemsets (HUIs) mined from the original database D and the sanitized database D', respectively. The loss utilities of the discovered HUIs before and after sanitization can be denoted as Itemset Utility Similarity (IUS), which is defined as:*

$$IUS = \frac{\sum_{X \in HUIs^{D'}} u(X)}{\sum_{X \in HUIs^{D}} u(X)}. \tag{7}$$

The conducted experiments of three algorithms under varied dataset sizes in two datasets are compared and shown in Fig. 5. From Fig. 5 , it can be seen that the proposed algorithm still has better results compared to the other algorithms since the minimum utility mechanism is applied to delete or decrease the utilities of items.

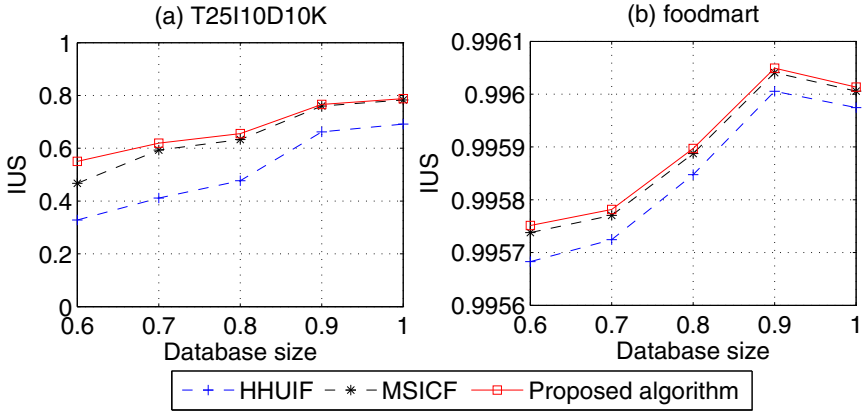


Fig. 5 Itemsets utility similarity (IUS) w.r.t. varied dataset sizes.

6 Conclusion

In this paper, an efficient algorithm is developed to efficiently delete the sensitive high utility itemsets or decrease the utilities of them based on minimum utility mechanism. Three novel criteria namely Database Structure Similarity (DSS), Database Utility Similarity (DUS), and Itemset Utility Similarity (IUS) are thus developed in this paper to clearly reveal the efficiency and effectiveness of the developed algorithms in PPUM. From the conducted experiments, the proposed algorithm generally has better results compared to the state-of-the-art algorithms in PPUM.

Acknowledgments This research was partially supported by the Tencent Project under grant CCF-TencentRAGR20140114, by the Shenzhen Peacock Project, China, under grant KQC201109020055A, and by the Natural Scientific Research Innovation Foundation in Harbin Institute of Technology under grant HIT.NSRIF.2014100.

References

1. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: The International Conference on Very Large Data Bases, pp. 487–499 (1994)
2. Agrawal, R., Srikant, R.: Privacy-preserving data mining. ACM SIGMOD Record **29**(2), 439–450 (2000)
3. Amiri, A.: Dare to share: Protecting sensitive knowledge with data sanitization. Decision Support Systems **43**(1), 181–191 (2007)
4. Atallah, M., Elmagarmid, A., Ibrahim, M., Bertino, E., Verykios, V.: Disclosure limitation of sensitive rules. In: The Workshop on Knowledge and Data Engineering Exchange, pp. 45–52 (1999)
5. Bertino, E., Fovino, I.N., Provenza, L.P.: A framework for evaluating privacy preserving data mining algorithms. Data Mining and Knowledge Discovery **11**(2), 121–154 (2005)
6. Chen, M.S., Han, J., Yu, P.S.: Data mining: An overview from a database perspective. IEEE Transactions on Knowledge and Data Engineering **8**(6), 866–883 (1996)
7. Dasseni, E., Verykios, V.S., Elmagarmid, A.K., Bertino, E.: Hiding association rules by using confidence and support. In: Moskowitz, I.S. (ed.) IH 2001. LNCS, vol. 2137, pp. 369–383. Springer, Heidelberg (2001)
8. Evfimievski, A., Srikant, R., Agrawal, R., Gehrke, J.: Fast algorithms for mining association rules in large databases. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 217–228 (2002)
9. Fournier-Viger, P., Gomariz, A., Gueniche, T., Soltani, A., Wu, C.W., Tseng, V.S.: SPMF: a Java Open-Source Pattern Mining Library. Journal of Machine Learning Research **15**, 3389–3393 (2014)
10. Giannotti, F., Lakshmanan, L.V.S., Monreale, A., Pedreschi, D., Wang, H.W.: Privacy-preserving mining of association rules from outsourced transaction databases. IEEE Systems Journal **7**(3), 385–395 (2012)
11. Goethals, B., Zaki, M.J.: Frequent itemset mining implementations repository (2012). http://fimi.ua.ac.be/data/
12. Han, J., Pei, J., Yin, Y., Mao, R.: Mining frequent patterns without candidate generation: A frequent-pattern tree approach. Data Mining and Knowledge Discovery **8**(1), 53–87 (2004)
13. Hong, T.P., Lin, C.W., Yang, K.T., Wang, S.L.: Using TF-IDF to hide sensitive itemsets. Applied Intelligence **38**(4), 502–510 (2013)
14. Li, X.B., Sarkar, S.: A tree-based data perturbation approach for privacy-preserving data mining. IEEE Transactions on Knowledge and Data Engineering **18**(9), 1278–1283 (2006)
15. Li, Y.C., Yeh, J.S., Chang, C.C.: MICF: An effective sanitization algorithm for hiding sensitive patterns on data mining. Advanced Engineering Informatics **21**(3), 269–280 (2007)
16. Lindell, Y., Pinkas, B.: Privacy preserving data mining. In: International Cryptology Conference on Advances in Cryptology, 36–54 (2000)
17. Liu, Y., Liao, W., Choudhary, A.K.: A two-phase algorithm for fast discovery of high utility itemsets. In: Ho, T.-B., Cheung, D., Liu, H. (eds.) PAKDD 2005. LNCS (LNAI), vol. 3518, pp. 689–695. Springer, Heidelberg (2005)
18. Lin, C.W., Hong, T.P., Lu, W.H.: An effective tree structure for mining high utility itemsets. Expert Systems with Applications **38**(6), 7419–7424 (2011)
19. Lin, C.W., Zhang, B., Yang, K.T., Hong, T.P.: Efficiently hiding sensitive itemsets with transaction deletion based on genetic algorithms. The Scientific World Journal **2014**, 1–13 (2014)

20. Lin, C.W., Hong, T.P., Wong, J.W., Lan, G.C., Lin, W.Y.: A GA-Based approach to hide sensitive high utility itemsets. The Scientific World Journal **2014**, 1–12 (2014)
21. Sun, X., Yu, P.S.: A border-based approach for hiding sensitive frequent itemsets. In: IEEE International Conference on Data Mining, pp. 27–30 (2005)
22. Verykios, V.S., Elmagarmid, A.K., Bertino, E., Saygin, Y., Dasseni, E.: Association rule hiding. IEEE Transactions on Knowledge and Data Engineering **16**(4), 434–447 (2004)
23. Verykios, V.S., Bertino, E., Fovino, I.N., Provenza, L.P., Saygin, Y., Theodoridis, Y.: State-of-the-art in privacy preserving data mining. ACM SIGMOD Record **33**(1), 50–57 (2004)
24. Yao, H., Hamilton, H.J., Butz, C.J.: A foundational approach to mining itemset utilities from databases. In: SIAM International Conference on Data Mining, pp. 482–486 (2004)
25. Yao, H., Hamilton, H.J.: Mining itemset utilities from transaction databases. Data and Knowledge Engineering **59**(3), 603–626 (2006)
26. Yeh, J.S., Hsu, P.C.: HHUIF and MSICF: Novel algorithms for privacy preserving utility mining. Expert Systems with Applications **37**(7), 4779–4786 (2010)
27. Yun, U., Kim, J.: A fast perturbation algorithm using tree structure for privacy preserving utility mining. Expert Systems with Applications **42**(3), 1149–1165 (2015)

Security Analysis of an Anonymous Authentication Scheme Based on Smart Cards and Biometrics for Multi-server Environments

Jeng-Shyang Pan, Raylin Tso, Mu-En Wu and Chien-Ming Chen

Abstract User authentication is an important technology for E-commerce, especially when it is done by using smart cards. Authentication schemes based on smart cards can guarantee that a user using the smart card is legal and has the authorization to access resources (eg., a bank account or a remote server) behind the smart card. Due to its usefulness, authentication schemes based on smart cards have been widely researched in recent years. In 2014, Choi introduced a security enhanced anonymous multi-server authenticated key agreement scheme using smart card and biometrics. Kuo et. al recently found that Choi's scheme is insecure against card losing attack and made an improvement to deal with the problem. However, in this paper, we will show that Kuo et. al's new scheme made the situation even worse. In their new scheme, any server having communicated with and received information from a card of a user can impersonate the user and enjoy the service (eg., on-line shopping) from the server on behalf of the original user without the card on-hand. We conduct a detailed analysis of flaws in their scheme in the hope that no similar mistakes are made in the future. An improved scheme is left as a future work.

J.-S. Pan
College of Information Science and Engineering,
Fujian University of Technology, Fujian, China
e-mail: jengshyangpan@gmail.com

R. Tso(✉)
Department of Computer Science, National Chengchi University, Taipei, Taiwan
e-mail: raylin@cs.nccu.edu.tw

M.-E. Wu
Department of Mathematics, Soochow University, Taipei, Taiwan
e-mail: mn@scu.edu.tw

C.-M. Chen
Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, China
e-mail: chienming.taiwan@gmail.com

T.T. Zin et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 388,
DOI: 10.1007/978-3-319-23207-2_7

Keywords Authentication · Biometric information · Multi-server environments · Security · Smart card

1 Introduction

With the rapid development of the Internet technology, more and more people have relied on Internet to access information, exchange knowledge, and process data in distributed network environments. Moreover, e-commerce such as on-line shopping, on-line transactions, on-line stock and on-line banking are more and more popular thanks to the Internet technology. To achieve e-commerce smoothly and successfully, network security as well as user authentication [2, 3, 7, 12] are indispensable. User authentication is an important technology to guarantee that only the legal users can access resources from a remote server. To achieve simplicity, efficiency and low-communications, the techniques of user authentication based on smart cards are becoming more and more important in today's network environments.

1.1 Motivations and Our Contributions

In recent years, due to the cryptographic capacity, low cost, and the portability, the smart card based authentication scheme is becoming more and more important and providing various functionality [4, 13, 14, 17, 18]. There are many remote user authentication protocols with smart card which have been proposed to improve security, efficiency, and functionality extensively by many scholars in recent years [6, 8, 9, 10, 19, 20]. Moreover, the compromise of user's identity would lead to the tracing of the previous network communications for the same user. To protect from the risk of ID-theft, the user anonymity property is required for the privacy protection of users [6, 20]. In 2008, Juang et al.'s [10] proposed a new password-authenticated key agreement protocol based on elliptic curve cryptosystems. Their scheme not only could provide identity protection but also construct the session key agreement and enhance efficiency by using elliptic curve cryptosystems. Unfortunately, Sun et al.'s proposed an improved scheme to overcome the weakness of Juang et al.'s, including inability of the password-changing and the session key problem [19]. Later, there are many password based authentications with smart card having been proposed to achieve the user anonymity [6, 9, 16, 20].

Due to the smart card usually does not support powerful computation capability, new authentication protocols with less calculation in the smart cards are required [10, 16, 19]. In addition, for security considerations and convenience from users perspective, many researchers proposed the smart card authentication scheme combined with biometric information to enhance overall security [5, 15, 21, 22]. It is pointed at [11] that biometric has the following five characteristics:

1. Universality: each entity should have its biometric trait.
2. Distinctiveness: any two of the entities should have different biometric features.
3. Permanence: biometric features do not change over time.

4. Collectability: biometric features are measurable with simple technical instruments.
5. Uniqueness: biometric is unique.

Biometric features including face, fingerprint, iris, hand geometry, palm print, voice pattern...etc and have come into limelight in recent years for personal authentication.

On the other hand, nowadays, ubiquitous computing has become very popular where multiple servers are involved in authenticating their users. In multi-server environments, single registration to a trusted registration center is the most important feature and any user could receive desired services from various service providers without repeating registration. Taking this advantage as a consideration, later on, many convenient authentication schemes with smart cards and biometric have been proposed for the multi-server environments [1, 5, 11, 22].

In 2014, Choi [5] introduced an enhanced anonymous authentication scheme using a smart card and biometric information for multi-server environment. Their scheme is aim to improve the weaknesses they discovered in Chuang and Chen 's scheme [1]. Recently, Kuo et al. investigated Chois's scheme [5] and found that their scheme is still insecure against smart card losing attack and has no forward secrecy when card lost. Kuo et al. consequently introduced an improved scheme [11] in 2015 and claimed that the new scheme can overcome all the weaknesses they discovered in Choi's scheme.

However, in this paper, we will show that Kuo et. al's new scheme is still insecure. Moreover, it made the situation even worse for card holders. In their new scheme, any server having communicated with and received information from a card of a user can impersonate the user and enjoy the service (eg., on-line shopping) from the server on behalf of the original user without the card on-hand. We will conduct a detailed analysis of flaws in their scheme in the hope that no similar mistakes are made in the future. An improved scheme is left as a future work.

The rest of the paper is organized as follows: in Section 2, we will review Choi's scheme [5] and show the weaknesses discovered in [11]. In Section 3, the new scheme introduced by Kuo et. al. in [11] is reviewed. Section 4 demonstrates the security weakness of Kuo et. al. 's scheme. Finally, conclusions are given in Section 5.

2 Review of Choi's Scheme and Its Security Weakness

We first review Choi's scheme and show the insecurity of the scheme. The presentation in this section follows that of Kuo et. al's paper [11]. Table 1 shows the notations used in Choi's scheme.

The scheme consists of three phases; registration phase, login phase and authentication phase.

Table 1 Notations of Choi's Scheme

Notations	Description
x	A secret value of the registration center
RC	The registration center
UID_i	The identification of user i
SID_i	The identification of server j
$AUID_i$	The anonymous identification of user i
$ASID_j$	The anonymous identification of server j
PW_i	The password of user i
BIO_i	The biometric information of user i
$h(.)$	A cryptographic one-way hash function
N_i	A random number
PSK	A secure and pre-shared key among RC and servers
$\|\|$	A string concatenation operation
$\oplus$	A string XOR operation

Registration Phase

- $U_i \rightarrow RC : \{UID_i, h(PW_i \oplus BIO_i)\}$
  User U_i with identity UID_i computes $h(PW_i \oplus BIO_i)$ and sends $\{UID_i, h(PW_i \oplus BIO_i)\}$ to RC via a secure channel.
- $RC \rightarrow U_i : \{UID_i, h(.), B_i, C_i, D_i, E_i, F_j\}$
  After receiving the message from U_i, RC with its secret value x computers A_i to E_i as follows:

1. $A_i = h(UID_i||x)$
2. $B_i = h^2(UID_i||x) = h(A_i)$
3. $C_i = h(PW_i \oplus BIO_i) \oplus B_i$
4. $D_i = PSK \oplus A_i$
5. $E_i = h(PSK) \oplus h(PW_i \oplus BIO_i)$
6. $F_i = [SID_1, SID_2, \ldots, SID_n]$

Then, RC stores $\{UID_i, h(.), B_i, C_i, D_i, E_i, F_i\}$ into the smart card and sends it back to U_i via a secure channel.

Login Phase

- $U_i \rightarrow SmartCard : \{UID_i, PW_i, BIO_i\}$
  U_i inserts the smart card and inputs his personal information $\{UID_i, PW_i, BIO_i\}$
- The smart card checks the identity UID_i and then computes $B' = h(PW_i \oplus BIO_i) \oplus C_i$ using the information received from the user U_i at the previous step.
- Accept the login request if UID_i is valid and $B' = B$. Otherwise, terminate the login phase.

Authentication Phase. When user U_i wants to access the resources of server S_j, the following steps are performed between the smart card and the server S_j.

- $SmartCard \rightarrow S_j : \{AUID_i, M_1, M_2, D_i, T_1\}$
  The smart card generates a new random number N_1 and computes $AUID_i$, M_1 and M_2 as follows:

  1. $M_1 = h(B_i) \oplus N_1 \oplus h(PSK)$
  2. $AUID_i = h(N_1||UID_i||SID_j)$
  3. $M_2 = h(AUID_i||SID_j||D_i||N_1||T_1)$, where T_1 is a timestamp.

  The smart card then sends $\{AUID_i, M_1, M_2, D_i, T_1\}$ to S_j through a public channel.
- $S_j \rightarrow SmartCard : \{ASID_j, M_3, M_4, T_2\}$
  S_j first checks the timestamp and proceeds the following steps if $T_2 - T_1 \leq \Delta T$. In this case, the timestamp T_1 is said to be valid.

  1. $A_i = D_i \oplus PSK$
  2. $N_1 = M_1 \oplus h^2(A_i) \oplus h(PSK)$
  3. $M_2' = h(AUID_i||SID_j||D_i||N_1||T_1)$

  S_j authenticates the smart card and recognizes it as legal if $M_2' = M_2$. In this case, S_j then generates a new random number N_2 and continues to calculate M_3, M_4, $ASID_J$ and SK_{ij} as follows:

  1. $M_3 = N_2 \oplus h^2(N_1)$
  2. $ASID_j = h(N_2||UID_i||SID_j)$
  3. $M_4 = h(AUID_i||ASID_j||D_i||N_1||T_2)$
  4. $SK_{ij} = h(N_1||N_2||AUID_i||ASID_j)$

  S_j then returns $\{ASID_j, M_3, M_4, T_2\}$ to the smart card of U_i.
- $SmartCard \rightarrow S_j : \{M_5, T_3\}$
  After receiving $\{ASID_j, M_3, M_4, T_2\}$ from S_j, the smart card checks the timestamp $T_3 - T_2 \leq \Delta T$. if timestamp is valid, the smart card computes $N_2' = M_3 \oplus h^2(N_1)$ and $M_4' = h(AUID_i||ASID_j||D_i||N_2'||T_2)$ and checks whether it is equal to M_4. If they are equal, the smart card continues the following steps:

  1. $SK_{ij} = h(N_1||N_2||AUID_i||ASID_j)$
  2. $M_5 = h(SK_{ij}||h(N_2)||T_3)$

  The smart card sends $\{M_5, T_3\}$ to S_j via a public channel.
- After receiving $\{M_5, T_3\}$, S_j checks the timestamp of T_3(ie., $T_4 - T_3 \leq \Delta T$ where T_4 is the current timestamp), if it is valid, then S_j computes $M_5' = h(SK_{ij}||h(N_2)||T_3)$. If $M_5' = M_5$, S_j verifies the smart card and the session key SK_{ij} is successfully established.

2.1 Security Weakness

Kuo et. al pointed out in [11] that Choi's scheme is insecure against card losing attack. If a card of a user is lost, then the session key will be compromised if the communication between the user and a server is eavesdropped by the attacker.

Assume that the attacker picks up U_i's smart card and gets $\{UID_i, h(.)B_i, C_i, D_i, E_i\}$ from the card. In addition, if the attacker has ever intercepted the communication between the card and a server S_j, he received $\{AUID_i, ASID_j, M_1, M_3\}$ from the communication. Then SK_{ij} is computed as follows:

- $h(PW_i \oplus BIO_i) = C_i \oplus B_i$
- $h(PSK) = E_i \oplus h(PW_i \oplus BIO_i)$
- $N_1 = M_1 \oplus B_i \oplus h(PSK)$
- $N_2 = M_3 \oplus h^2(N_1)$
- $SK_{ij} = h(N_1||N_2||AUID_i||ASID_j)$

This proves that Choi's scheme does not provide forward security when card is lost.

Furthermore, Kuo et. al. also showed that the anonymity is not preserved and the scheme is suffered from the impersonation attack once the card is lost. So, it is concluded that Choi's scheme is vulnerable against card losing attack.

3 Kuo et. al. 's Scheme Revisited

Based on Choi's scheme, in 2015, Kuo et. al. proposed an enhanced anonymous authentication key agreement scheme [11] . The new scheme is aim to overcome all of the security issues of Choi's Scheme. Their new scheme consists of three main phases: registration phase, login phase and authentication phase. The notations used here is the same as those in Choi's scheme.

Registration Phase

- $U_i \rightarrow RC : \{UID_i, PW_i, BIO_i\}$
  User U_i with identity UID_i sends $\{UID_i, PW_i, BIO_i\}$ to RC via a secure channel.
- $RC \rightarrow U_i : \{h(.), C_i, D_i, E_i\}$
  After receiving the message from U_i, RC with its secret value x computers A_i to E_i as follows:

1. $A_i = h(UID_i||x)$
2. $B_i = h^2(UID_i||x) = h(A_i)$
3. $C_i = h(PW_i \oplus BIO_i \oplus UID_i)$
4. $D_i = PSK \oplus A_i$
5. $E_i = UID_i \oplus B_i$

Then, RC stores $\{h(.), C_i, D_i, E_i\}$ into the smart card and sends it back to U_i via a secure channel.

Login Phase

- $U_i \rightarrow SmartCard : \{UID_i, PW_i, BIO_i\}$
  U_i inserts the smart card and inputs his personal information $\{UID_i, PW_i, BIO_i\}$
- The smart card computes $C_i' = h(PW_i \oplus BIO_i \oplus UID_i)$ using the information received from the user U_i at the previous step.
- Accept the login request if $C_i' = C_i$. Otherwise, terminate the login phase.

Authentication Phase. When user U_i wants to access the resources of server S_j, the following steps are performed between the smart card and the server S_j.

- $SmartCard \rightarrow S_j : \{AUID_i, M_1, M_2, D_i, T_1\}$
  The smart card generates a new random number N_1 and computes $AUID_i$, M_1 and M_2 as follows:

  1. $B_i = UID_i \oplus E_i$
  2. $M_1 = h(B_i) \oplus N_1$
  3. $AUID_i = h(N_1) \oplus UID_i \oplus BIO_i$
  4. $M_2 = h(AUID_i||SID_j||D_i||N_1||T_1)$, where T_1 is a timestamp.

  The smart card then sends $\{AUID_i, M_1, M_2, D_i, T_1\}$ to S_j through a public channel.
- $S_j \rightarrow SmartCard : \{SID_j, M_3, M_4, T_2\}$
  S_j first checks the timestamp and proceeds the following steps if $T_2 - T_1 \leq \Delta T$. In this case, the timestamp T_1 is said to be valid.

  1. $A_i = D_i \oplus PSK$
  2. $N_1 = M_1 \oplus h^2(A_i)$
  3. $M_2' = h(AUID_i||SID_j||D_i||N_1||T_1)$

  S_j authenticates the smart card and recognizes it as legal if $M_2' = M_2$. In this case, S_j then generates a new random number N_2 and continues to calculate M_3, M_4 and SK_{ij} as follows:

  1. $M_3 = N_2 \oplus h^2(N_1)$
  2. $M_4 = h(AUID_i||SID_j||N_2)$
  3. $SK_{ij} = h(N_1||N_2)$

  S_j then returns $\{SID_j, M_3, M_4, T_2\}$ to the smart card of U_i.
- $SmartCard \rightarrow S_j : \{M_5, T_3\}$ After receiving $\{SID_j, M_3, M_4, T_2\}$ from S_j, the smart card checks the timestamp $T_3 - T_2 \leq \Delta T$. if timestamp is valid, the smart card computes $N_2' = M_3 \oplus h^2(N_1)$ and $M_4' = h(AUID_i||SID_j||N_2')$ and checks whether it is equal to M_4. If they are equal, the smart card continues the following steps:

1. $SK_{ij} = h(N_1||N_2)$
2. $M_5 = h(SK_{ij}||h(N_2))$

The smart card sends $\{M_5, T_3\}$ to S_j via a public channel.

- After receiving $\{M_5, T_3\}$, S_j checks the timestamp of T_3 (ie., $T_4 - T_3 \leq \Delta T$ where T_4 is the current timestamp), if it is valid, then S_j computes $M_5' = h(SK_{ij}||h(N_2))$. If $M_5' = M_5$, S_j verifies the smart card and the session key SK_{ij} is successfully established.

4 Security Analysis on Kuo et. al.'s Scheme

The first security concern is about their registration phase. To avoid card losing attack, in Kuo et. al.'s scheme, PW_i and BIO_i are sent in plaintext form to RC (ie., the registration center). This means that RC stores and possesses all the passwords and biometric informations of users having ever registered in the system. We all know that passwords and biometric informations are very sensitive, especially our biometric informations. Biometric information should not leak to anyone since it does not change over time and is unique to us. Once it is disclosed, it may be abused by others. If the RC system is vulnerable, attackers may hack the system and retrieve our biometric informations. The administrator of RC may also abuse our personal information if he/she is not so trustworthy. Consequently, it cannot be recommended to allow anyone other than ourself to posses our biometric information.

Secondly, we will show that Kuo et. al.'s scheme is vulnerable to insider attacks. Here insider means a server having communicated with a user and exchanged information with the user before. In this attack, the server with only public information can impersonate the user to communicate with other servers and finally get authenticated and exchanged keys by the server. Most importantly, this attack uses only public information and does not need a smart card of a user on hand. This means that a user may not aware of his/her card being abused since the card is not lost.

Assume a server S_j has ever communicated with a user U_i, we describe the attack procedure in detail in the following.

Pre-computation Phase

- S_j has the following information since it has communicated with U_i before.
  - $M_1 = h(B_i) \oplus N_1$
  - $AUID_i = h(N_1) \oplus UID_i \oplus BIO_i$
  - $M_2 = h(AUID_i||SID_j||D_i||N_1||T_1)$
  - D_i

  These are received at the first step of the authentication phase

- S_j recovers N_1 from
  - $A_i = D_i \oplus PSK$
  - $N_1 = M_1 \oplus h^2(A_i) = M_1 \oplus h(B_i)$

Impersonation Phase. To impersonate U_i and to cheat a server S_k, S_j has to send valid information $\{AUID_i', M_1', M_2', D_i, T_1'\}$ to S_k for authentication. These can be done as follows:

- Pick N_1' at random.
- $M_1' = M_1 \oplus N_1 \oplus N_1'$
- $AUID_i' = AUID_i \oplus h(N_1) \oplus h(N_1')$
- $M_2' = h(AUID_i'||SID_k||D_i||N_1'||T_1')$

Then, S_j can impersonate U_i and send $\{AUID_i', M_1', M_2', D_i, T_1'\}$ to S_k.

S_k will authenticate the attacker S_j as a valid user via checking M_2' with the value he computed (following step two of the authentication phase of Kuo et. al.'s scheme). After than, S_k will compute $M_3' = N_2' \oplus h^2(N_1')$, $M_4' = h(AUID_i||SID_k||N_2')$ and $SK_{ik} = h(N_1'||N_2')$. $\{M_3', M_4'\}$ are returned back to the fake user (i.e., S_j).

Session Key Discovery Phase. From M_3', the session key SK_{ik} can be computed as follows:

- $N_2' = M_3' \oplus h^2(N_1')$
- $SK_{ik} = h(N_1'||N_2')$

M_5' can be computed accordingly so, at the end, S_j successfully cheated the server S_k and can impersonate U_i to enjoy the service provided from S_k using the identity U_i.

5 Conclusion

Recently, Kuo et al. investigated Chois's scheme and found that their scheme is insecure against smart card losing attack and has no forward secrecy whey card lost. Kuo et al. consequently introduced an improved scheme in 2015 and claimed that the new scheme can overcome all the weaknesses they discovered in Choi's scheme. In this paper, we showed that Kuo et. al's new scheme is still insecure. In their new scheme, any server having communicated with and received information from a card of a user can impersonate the user and make on-line shopping without the card on-hand. We conducted a detailed analysis of flaws in their scheme in the hope that no similar mistakes are made in the future. An improved scheme is left as a future work.

Acknowledgments The work of Raylin Tso was supported in part by the Ministry of Science and Technology, Taiwan, R.O.C., under Grant MOST 103-2221-E-004-009. The work of Mu-En Wu was supported in part by the Ministry of Science and Technology, Taiwan, R.O.C., under Grant MOST 103-2218-E-031-001. The work of Chien-Ming Chen was supported in part by the Project NSFC (National Natural Science Foundation of China) under Grant number 61402135.

References

1. Chuang, M.C., Chen, M.C.: An anonymous multi-server authenticated key agreement scheme based on trust computing using smart cards and biometrics. Expert Systems with Applications **41**(4), 1411–1418 (2014)
2. Chen, C.-M., Ku, W.-C.: Stolen-verifier attack on two new strong-password authentication protocols. IEICE Transactions on Communications **85**(11), 2519–2521 (2002)
3. Chen, C.-M., Wang, K.H., Wu, T.Y., Pan, J.S., Sun, H.M.: A Scalable Transitive Human-Verifiable Authentication Protocol for Mobile Devices. IEEE Transactions on Information Forensics and Security **8**(8), 1318–1330 (2013)
4. Chien, H.Y., Jan, J.K., Tseng, Y.M.: An efficient and practical solution to remote authentication: Smart Card. Computer & Security **21**, 372–375 (2002)
5. Choi, Y., Nam, J., Lee, D., Kim, J., Jung, J., Won, D.: Security enhanced anonymous multi-server authenticated key agreement scheme using smart card and biometrics. The Scientific World Journal **2014**, Article 281305 (2014)
6. Das, M.L., Saxena, A., Gulati, V.P.: A dynamic ID-based remote user authentication scheme. IEEE Transactions on Consumer Electronics **50**(2), 629–631 (2004)
7. Farash, M.S., Attari, M.A.: An efficient and provably secure three-party password-based authenticated key exchange protocol based on Chebyshev chaotic maps. Nonlinear Dynamics **77**, 399–411 (2014)
8. Hwang, M.S., Chong, S.K., Chen, T.Y.: DoS resistant ID-based password authentication scheme using smart cards. Journal of Systems and Software **83**, 163–172 (2010)
9. He, D.J., Ma, M., Zhang, Y., Chen, C., Bu, J.J.: A strong user authentication scheme with smart cards for wireless communications. Computer Communication **34**, 367–374 (2011)
10. Juang, W.S., Chen, S.T., Liaw, H.T.: Robust and efficient password-authenticated key agreement using smart card. IEEE Transactions on Industrial Electronics **5**, 2551–2556 (2008)
11. Kuo, W.C., Wei, H.J., Chen, Y.H., Chen, J.C.: An enhanced secure anonymous authentication scheme based on smart cards and biometrics for multi-server environments. In: Proc. of The 10th Asia Joint Conference on Information Security (AsiaJCIS 2015) (2015)
12. Ku, W.-C., Chen, C.-M., Lee, H.-L.: Cryptanalysis of a variant of Peyravian-Zunic's password authentication scheme. IEICE Transactions on Communications **86**(5), 1682–1684 (2003)
13. Lee, N.Y., Chiu, Y.C.: Improved remote authentication scheme with smart card. Computer Standards & Interfaces **27**, 177–180 (2005)
14. Lee, S.W., Kim, H.S., Yoo, K.Y.: Improvement of Chien et al.s remote user authentication scheme using smart cards. Computer standards & Interfaces **27**(2), 181–183 (2005)
15. Liu, M., Shieh, W.G.: On the security of Yoon and Yoo's biometrics remote user authentication scheme. WSEAS Transactions on Information Science and Applications **11**, 94–103 (2014)
16. Song, R.: Advanced smart card based password authentication protocol. Computer Standards & Interfaces **32**, 321–325 (2010)
17. Sun, H.M.: An efficient remote user authentication scheme using smart cards. IEEE Transactions on Consumer Electronics **46**, 958–961 (2000)
18. Sun, H.M., Hung, C.F., Chen, C.M.: An improved digital rights management system based on smart cards. In: Proc. of Digital EcoSystems and Technologies Conference (DEST 2007) (2007)

19. Sun, D.Z., Huai, J.P., Sun, J.Z., Li, J.X., Zhang, J.W., Feng, Z.Y.: Improvements of Juang et al.s password-authenticated key agreement scheme using smart cards. IEEE Transactions on Industrial Electronics **56**, 2284–2291 (2009)
20. Wang, Y., Liu, J., Xiao, F., Dan, J.: A more efficient and secure dynamic ID-based remote user authentication scheme. Computer Communications **32**(4), 583–585 (2009)
21. Wu, J., Zhu, W.T., Feng, D.G.: Improvement of a fingerprint-based remote user authentication scheme. International Journal of Security and its Applications **2**(3), 208 (2008)
22. Yang, D., Yang, B.: A biometric password-based multi-server authentication scheme with smart card. IEEE International Conference on Computer Design and Applications **5**, 554–559 (2010)

A Modeling Method of Virtual Terrain Environment

Lian-Lei Lin, Ling-Yu Li and Xin-Yi Song

Abstract Terrain data is the most commonly used data in virtual test. In various applications, terrain data's type and representation are also different, so we often need to convert terrain data according to the application demand. Aiming at this problem, this paper proposes a new terrain environment modeling method based on SEDRIS (Synthetic Environment Data Representation and Interchange Specification), which we used to represent and exchange the terrain data. Firstly, the form and characteristics of all kinds of terrain data are analyzed, and then the appropriate SEDRIS standard DRM class and the EDCS dictionary are selected to represent the original data, and the STF format terrain environment data with SEDRIS standard is generated. The research of this paper can enhance the normative representation and conversion efficiency of terrain environment data, and realize the terrain data sharing and reuse.

Keywords Virtual test · Terrain environment · SEDRIS

1 Introduction

As the most complicated and widely-used component of synthetic nature environment, terrain environment has close relation to all kinds of modeling and simulation system[1], such as action model, maneuvering model with direct information interchange and all kinds of environment model influenced by terrain etc.. Along with the development of simulation technology, many simulation field exploited terrain environment database which can meet their demand[2][3][4]. Under the condition, with no effective interchange mechanism, all kinds of terrain environment database can only use the particular way, point to point between

L.-L. Lin(✉) · L.-Y. Li · X.-Y. Song
Department of Automatic Testing and Control, Harbin Institute of Technology,
No.2 Yi-Kuang Street, Nangang District, Harbin 150080, China
e-mail: linlianlei@hit.edu.cn

T.T. Zin et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 388,
DOI: 10.1007/978-3-319-23207-2_8

systems, to exchange, low conversion efficiency and high maintenance cost. Therefore, we need use efficiency method to representation and exchange terrain environment data.

SEDRIS can realize environment data representation, intact, clearness, unambiguous, multi-state, and can realize the sharing, interchange, reusing of environment. After researching SEDRIS specification for many years internally [5][6], there are some descriptions of naval battlefield environment[7], atmosphere environment[8] and battlefield situation information[9] based on SEDRIS. Due to the complexity of terrain data, the presentation of terrain data based on SEDRIS has not been reported internally. Firstly, we analyzed the types of terrain data's all elements. Secondly, we also selected applicable SEDRIS model to express terrain environment data according to their own characteristic, and put forward the method of exchanging terrain environment data based on SEDRIS APIs, so that we construct synthetic environment database with STF (SEDRIS Transmittal Formal) format.

2 Analysis of Terrain Environment Element

Terrain environment data is complex, mainly divided into four types: altitude data, texture data, culture characteristic data and 3D model file, and this four type data describe terrain data from different aspects. It needs to establish different resolution models in a simulation system, including different resolution model in one simulation application. So that establishing synthetic terrain environment database need consider the type, property, resolution, range, characteristics changing with the time and the space etc., of which ,all kinds of terrain environment element.

2.1 Altitude Data

Altitude data is the digitalize representation for terrain surface shape, dividing into regular grid and irregular grid[3]. This paper chooses the altitude data of regular grid, and the vertex is evenly distributed. The common altitude data sources are USGS DEM and USGS DTED. USGS DEM file is constituted by logic record A, B, C. Type A, file header record, mainly records the information relation to altitude data; type B is profile data, including profile header data (the maximum and minimum of profile data) and profile object data; type C, precision information, is always omitted. USGS DTED uses the area, covered by 1 degree in longitude X 1 degree in latitude, as one file unit, including file header and data, and terrain point is decided by latitude and longitude grid.

For regular grid altitude data, the best method to establish multi-resolution model is to build regular tree structure based on terrain altitude point, regular tree structure has mature and high efficient node traversal algorithm [11]. This paper use quad-tree structure, every node in the tree covers one rectangular area in terrain, and root node covers all terrain area. The area covered by child node is a quarter of the area covered by father node, but the resolution is twice. At the same time, we divided altitude data into several blocks for more quickly retrieve.

2.2 Texture Data

Texture data is the image that used to show ground feature, generally satellite picture, for corresponding to altitude data. Texture data always choose satellite picture, so that need to clip because its big size. Meanwhile, texture picture may be stitched by many pictures in sequence. Choosing GDAL, who can operate all gridded geography data format, to read out, write in, transition, and dispose.

2.3 Cultural Characteristic Data

Cultural characteristic data is the digital vector file formed by various kind of map element which is vector. The vector descriptions of nature or humanistic characteristic is mainly about control point, settlement place and working condition construction, traffic and subsidiary facilities, water system and subsidiary facilities, boundary, physiognomy, soil texture, vegetation etc., about the place relationship and relevant property relationship between different elements, among the most popular is ERIS Shapefile file. Shapefile use point, line, polygon to store the shape of element, but cannot store topology relationship.

2.4 3D Model File

3D model is the geometry representation for terrain surface and the architecture, vegetation and other who are embedded in the terrain surface[4]. Most simulation systems support the 3D model format, OpenFlight file, created by Creator Terrain Studio software. OpenFlight file describe 3D object by geometry hierarchy structure and node property such as database header node, group, object, surface, polygon and so on. The simplest method, saving OpenFlight file into STF format database, is to find the common part between OpenFlight and SEDRIS data representation model. There are many nodes in OpenFlight structure can correspond with the class in SEDRIS.

3 Terrain Environment Data Representation Based on SEDRIS

SEDRIS includes two aspects of function target, which are data representation and data exchange. SEDRIS is mainly achieved by five core technique subassembly: Data Representation Model (DRM), Spatial Reference Model (SRM), Environment Data Coding Specification (EDCS), SEDRIS Interface Specification (API), SEDRIS Transmittal Formal (STF). Among them, SDRM, EDCS, SRM are used to realize the environmental data representation, and API and STF are used to realize data interchange. SDRM is the core of SEDRIS technology, using the object-oriented thought and method, provides a complete data representation and exchange mechanism for the modeling and the natural environment, and can

clearly describe the environmental data and environmental data logical relationship and relevance. SRM realizes the representation space position and coordinate transformation between different coordinate systems in spatial reference in different coordinate systems. The EDCS defines the semantics of SEDRIS, and provides a standard semantic interpretation for all objects attributes and data. The EDCS dictionary is the core of EDCS specification, including the classification/feature of SEDRIS, attribute codes and state codes. STF defines a middle database format and platform independent environment, support all SDRM data description information format, and realizes the cross platform data exchange environment. SEDRIS achieve data access to by hierarchical APIs.

Figure 1 provides all kinds of simulation application system based on SEDRIS terrain environment data by the data representation model of SEDRIS. All emulator, sensor, map creating system gain terrain environment data needed by SEDRIS' standard port, then change it into the data format they need. This kind of structure makes the standardability and availability of environment data representation better.

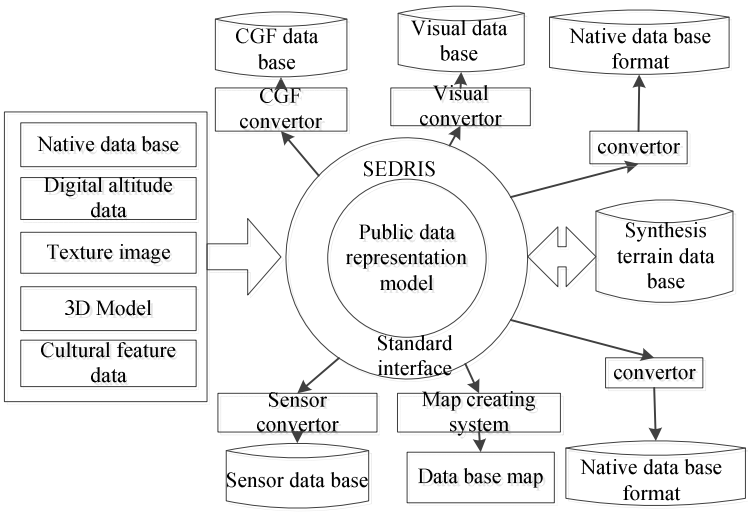


Fig. 1 Simulation system based on SEDRIS terrain environment database

3.1 Data Representation Model DRM

For all kinds of terrain environment data, there are four kinds of data representation model correspondingly, respectively raster data model, gridded data model, vector data model and polygonal data model[12][13].

(1) Raster Data Model: dividing terrain into regular grid array, defined by row-column, and the code it contains is property type or magnitude of that point/node. It is usually the striograph and scanogram information of texture data. Raster data

is saved in Image Object Class, all Image Object Class saved in Image Library Class, mapping into the particular spatial position by Image Anchor Class, the same image can be used into texture data by many terrain model, and can be related to altitude data or characteristic data by Image Mapping Function. Figure 3 shows the representation model of material.

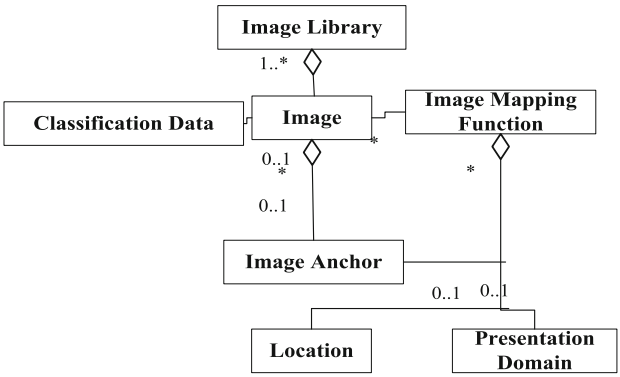


Fig. 2 Texture data representation model

(2) Gridded Data Model: storing environment data measured value or estimated value who is relation to particular spatial position. Terrain altitude, depth of sea and atmosphere data are described by 2D or 3D gridded structure usually. Gridded data is stored in Property Grid objects Class, all Property Grid Class have one spatial coordinate axis at least, saved in Data Table Library Class, which can be related to spatial position. Corresponding data representation model is as Figure 4. LOD Related Geometry Class realizes the multi-resolution setting of terrain data, data relating to corresponding resolution level by Index LOD Data Class. At the same time, saving data into blocks with Spatial Index Related Geometry Class, relating by one block saving in Index LOD Data and Property Grid Class. If the data is time varying, it can be saved under time point by using Time Related Geometry. And the time point, described by Time constraints Data Class, is associated with corresponding LOD Related Geometry.

(3) Vector Data Model: the set of terrain feature abstraction, mostly point, line, face, object and describing the topological between each other. It always is used to describe coastline, roadway, vegetation, soil, as well as building, factory, and other cultural feature data. Vector data is saved into Feature objects Class, various Feature objects classes can be related by texture data and altitude data. All characteristic data are divided into simple geometric feature (Union of Features Class), and the corresponding boundary of topological relations (Perimeter Related Feature Topology Class). Geometry features are divided into Point Feature Class, Liner Feature Class and Areal Feature Class.

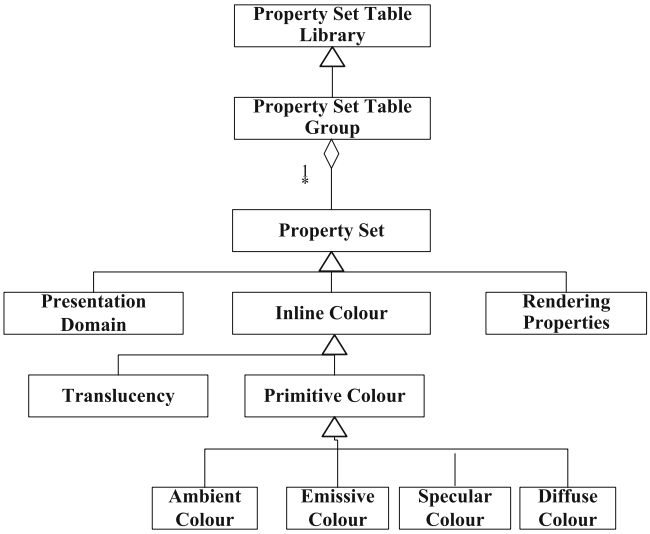


Fig. 3 Material data representation model

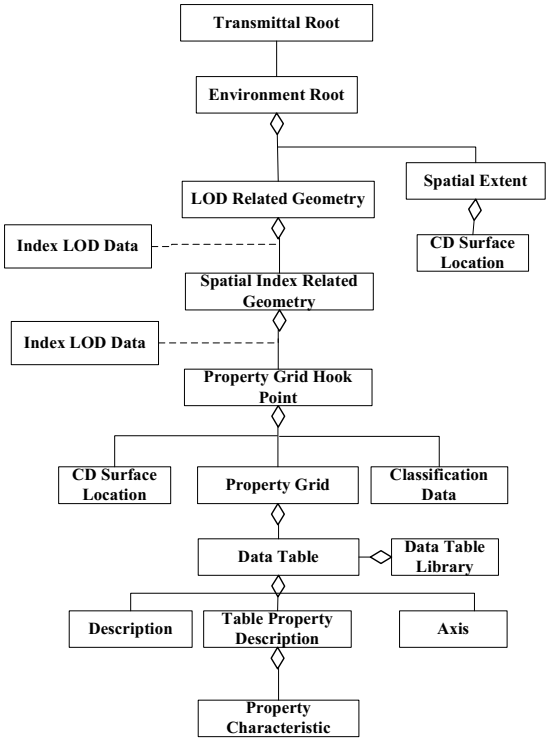


Fig. 4 Altitude data representation model

(4) Polygonal Data Model: the group of essential geometry element (point, line, polygon). It is usually used to describe terrain surface and the geometry embedded into the terrain surface. Table 1 is the comparison of SEDRIS data representation model and OpenFlight model [14]. At the same time, we can define data extension plugin and extension instrument plugin in Multigen Creator. Read and display SEDRIS format terrain environment data by OpenFlight API and SEDRIS API.

Table 1 The data representation comparison of OpenFlight and SEDRIS

Description	OpenFlight	SEDRIS
Header record	Header node	Geometry Model
Group record	Group node	Union of geometry hierarchy
Mesh record	Geometry node	Union of primitive features
LOD record	LOD node	LOD related geometry
Switch record	Switch – node	State related geometry
Object	Object node	Union of primitive geometry
Face record	Polygon node	Polygon
Vertex record	Vertex node	Vertex

Polygonal data is saved as Geometry objects Class, and different Geometry objects classes can be managed by texture data and altitude data. For 3D model file, it is saved by polygonal data model. Dividing 3D model into essential geometries, such as Point, Line, Arc, polygon, Ellipse, Volume Object etc.. Every point has its own texture settings, and can set properties like viewpoint and so on.

3.2 Data Coding EDCS

EDCS provides standard semantic interpretation for all object, property and data. For example, TREE, need to be described from TREE_BLOWDOWN, TREE_LINE, TREE_TRACT, and other aspects. The corresponding property information includes TREE_CANOPY_LEVEL_COUNT, TREE_CANOPY_BOTTOM_HEIGHT, TREE_TYPE, TREE_COUNT, TREE_SPACING.

4 Terrain Environment Data Based on SEDRIS

The terrain environment data described by the SDRM above can be written into STF by Write API, the port standard provided by SEDRIS. Firstly, create transmission. For all the objects presented by SDRM, Transmittal is instantiated by CreateObject, and assignment and setting corresponding properties by PutFields.

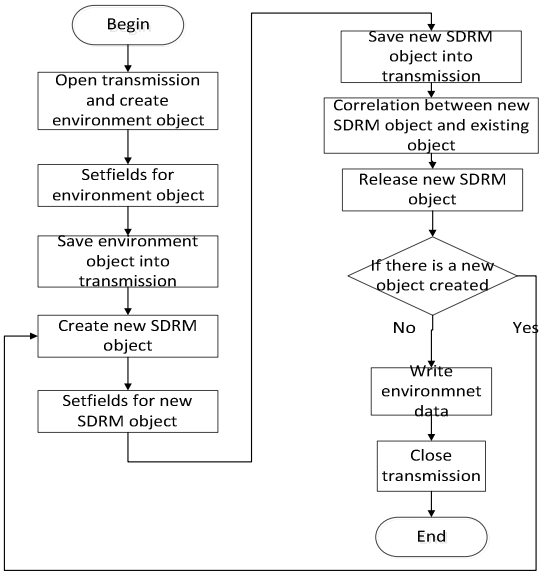


Fig. 5 Write API flow

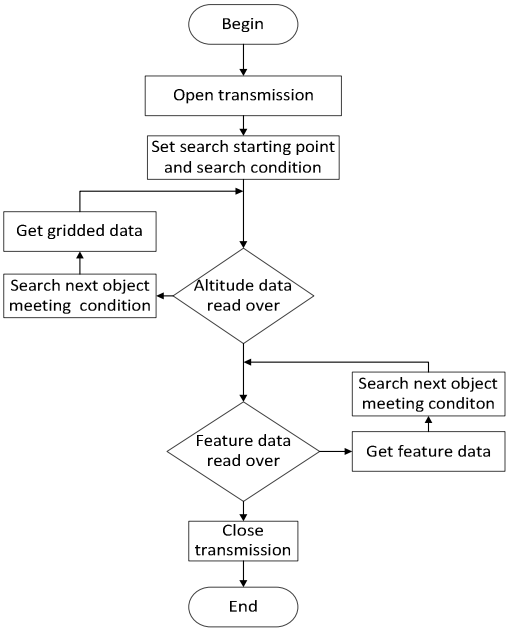


Fig. 6 Read API flow

Secondly, add the relationship between objects. Adding aggregation relationship by Addcomponent, and adding association by AddAssociate. Finally, adding it into Transmittal by setRootObject. Reading terrain environment data from STF based on Read API, can get specific classes by visiting aggregation node or component node directly, as well as, we can traverse STF database by iterating seIterator class and seSearchIterator class, then getting the objects. We choose seIterator for the object who need traverse only one layer, seSearchIterator for who need traverse multi-layer. Starting from one object, we do iteration traverse for other objects. Figure 5 and 6 gives us the flow chart of Write API and Read API.

5 Conclusions

SEDRIS technique accelerates interoperability between simulations and the reuse of environment data, as an effective data representation and interchange mechanism. This paper researches the terrain environment of synthesis natural environment, and introduces the elements and the creation of synthesis environment database according with SEDRIS standard detailed. By using SEDRIS API, other simulation systems can gain terrain environment data from synthesis terrain environment database, to build native terrain database, thus improving the utilization efficiency of terrain environment data.

Acknowledgment. This work is supported by National Science Foundation of China under Grant no. 61201305.

References

1. Gang, G.: Research on Synthetic Natural Environment Modeling and Simulation. National University of Defense Technology, October 2004
2. Xie, K., Huang, X., Wen, W.: Reserch on Environmental Data Reuse Method Based on Interchange Mechanism. Computer Applications and Software **29**(7), 53–55 (2012)
3. Weibing, M.: Research on the Synthetic Terrestrial Environment for Ground-to-air Missile Test Range Simulation. National University of Defense Technology (November 2003)
4. Jia, G., Wang, X.: An Effective Approach for Selection of Terrain Modeling Method. IEEE Geoscience and Remote Sensing Letters 10(4), July 2003
5. Yang, S., Zhan, S., Fei, Q.: Environment Data Representation and Interchange Using SEDRIS. Computer Engineering **28**(12), 71–74 (2002)
6. Haozhe, L., Jun, Z., Lin, L., Guohui, L.: Environment Data Representation Method for Synthetic Battle field Based on SEDRIS. Computer and Modernization, 17−19, July 2008
7. Xie, K., Zhao, G., Jiang, X., Huang, X., Wen, W.: A SEDRIS-BASED Description Method for Sea Battlefield Natural Environment. Computer Applications and Software **29**(6), 196–198 (2012)

8. Fang, Y.: Development of Atmospheric Environment Resource of Test and Training Architecture. Harbin Institute of Technology Instrument Science and Technology, pp. 1–50 (2012)
9. Lede, T., Jianwei, Q., Jiarun, W.: Battlefield Situation Information Describing Method Based on SEDRIS. Computer and Modernization **1**, 231–234 (2014)
10. Wang, J., Lin, J.: Multi-Resolution Model of Synthetic Natural Environment **10**, 176–180, July 2008
11. Hu, G., Zhang, G., Liu, W.: Modeling of Synthetic Terrain Environment and Implementation of Database. Computer Simulation **23**(7), 173–177 (2006)
12. Kruchhans, M.: ISO and OGC compliant database technology for the development of simulation object databases. In: Proceeding of the 2012 Winter Simulation Conference (2012)
13. Advanced Application of the DRM–Terrain. SEDRIS Document. http://www.sedris.org/paper/
14. Xu, M.: Research on Multi-resolution Representation and Visualization for Terrain Model in Virtual Reality. Wuhan University, December 2003
15. Environmental Data Coding Specification, International Standard ISO/IEC 18025:2005(E)

Method of Founding Focusing Matrix for Two-Dimensional Wideband Signals

Jiaqi Zhen, Zhifang Wang, Lipeng Gao, Hongyuan Gao and Ruihai Yang

Abstract The super-resolution direction finding for wideband signals usually requires preliminary direction of arrival(DOA) estimation, whether it is accurate or not will play an important part to the final result. In order to avoid the process, paper proposed a method of founding focusing matrix for two-dimensional wideband signals without preliminary DOA estimation, it is founded on Robust coherent signal subspace method (R-CSSM), the results achieved has a preferable robustness and higher precise than conventional Rotational signal subspace(RSS) method, wherever, there are no special requirements for the array manifold, computer simulations proved the effective performance of the method.

Keywords Direction of arrival estimation · Wideband signal · Focusing matrix · Robust coherent signal subspace method

1 Introduction

The spatial spectrum estimation super-resolution algorithms are widely used in radar, sonar and mobile communication in recent years, for example, multiple signal classification (MUSIC) [1] and estimation of signal parameters via rotational invariance techniques (ESPRIT) [2] are two representative methods of them, but they are only adapt to narrowband signals.

DOA estimation methods for wideband signals can be classified into two groups: Incoherent signal subspace method (ISSM) [3] and Coherent signal

J. Zhen(✉) · Z. Wang · R. Yang
Electronic Engineering, Heilongjiang University, Harbin, Heilongjiang, China
e-mail: zhenjiaqi2011@163.com, xiaofang_hq@126.com, 1176304741@qq.com

L. Gao · H. Gao
College of Information and Communication Engineering, Harbin Engineering University, Harbin, Heilongjiang, China
e-mail: 156083143@qq.com, yingruxiansheng@163.com

T.T. Zin et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 388,
DOI: 10.1007/978-3-319-23207-2_9

subspace method (CSSM) [4-7], the former needs to eigen-decompose to the covariance of every frequency, then calculate their spectrum function, so its calculation is very complicated, and the performance of some of the frequency is poor, this method needs a higher signal to noise ratio (SNR). The CSSM is proposed to the problem, it uses the idea of focusing, array manifold matrix of every frequency is aligned to the reference point, such as RSS [5], SST [6] and TCT [7], reducing the amount of calculation relative to ISSM, then many improving algorithms are proposed based on CSSM, Sellone [8] proposed Robust coherent signal subspace method based on RSS and SST methods, optimizing the freedom degree further, and it avoids pre-estimated of DOA. Feng [9] used Focusing Khatri-Rao (FKR) subspace method to change the covariance matrix to be a multiple dimensional matrix, estimated DOA of wideband signals by founding Khatri-Rao subspace, its precision is higher than classical CSM, but the calculation is still complexity. Huang kesheng [10] founded the focusing matrix by Jacobi-Anger spreading direction matrix, then acquired the Krylov subspace of array covariance matrix by multi-stage weiner filter (MSWF), reducing the amount of calculation greatly. Palanisamy [11] and Zhang jin[12] estimated DOA of wideband signals by propagator method, eliminated the colored noise by space difference technology.

The paper proposed a new method of focusing matrix for two-dimensional wideband signals based on robust coherent sub-space method, it uses iterative process. First, a group of focusing matrices with robustness to the whole angle space are generated, so the rough DOA can be acquired, shrink the searching area, then another group of focusing matrices with robustness to the angle space after shrinking are obtained, repeat the iteration process many times, the DOA of the higher precision can be calculated.

2 Array Signal Model

It is shown in Fig.1, the setting of the source detection problem is stated as followed: assume that N far-field wideband signals impinge on M-element ($N<M$) arbitrary placed plane array from distinct directions $(\theta_1,\varphi_1),\cdots,(\theta_N,\varphi_N)$, θ_i and φ_i is the

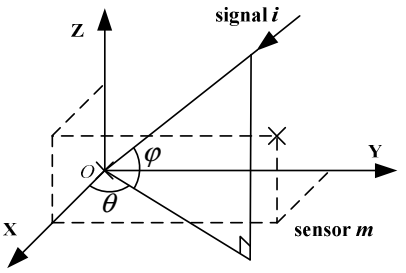


Fig. 1 Signal Model

azimuth and elevation of the ith signal, the coordinates of the mth sensor is (x_m, y_m) $(m = 1, 2, \cdots, M)$, the output of the ith sensor can be expressed:

$$x_m(t) = \sum_{i=1}^{N} s_i(t + \tau_{mi}) + n_m(t) \quad (m = 1, 2, \cdots, M) \tag{1}$$

where $\tau_{mi} = \dfrac{x_m \cos\theta_i \cos\varphi_i + y_m \sin\theta_i \cos\varphi_i}{c}$, $n_m(t)$ is the temporally and spatially white Gaussian noise of mth sensor, suppose the observation time of data collection is ΔT, the initial sampling frequency of incident signal on each sensor is f_s, the sampling times is $K = \Delta T f_s$, equation (1) can be transformed by Discrete Fourier Transform(DFT):

$$X_m(f) = \sum_{i=1}^{N} S_i(f) \exp(-\mathrm{j}2\pi f \tau_{mi}) + N_m(f) \tag{2}$$

then the output of the array can be decomposed into some narrowband parts by filter bank, that is

$$\boldsymbol{X}(f_i) = \boldsymbol{A}(f_i, \theta_i, \varphi_i)\boldsymbol{S}(f_i) + \boldsymbol{N}(f_i) \quad i = 1, 2, \cdots, J \tag{3}$$

where

$$f_i = \frac{i}{K} f_s \tag{4}$$

3 Principle of the New Method

3.1 RSS Method

As there is no focusing loss in the process when the focusing matrix is unitary, the constraint of this condition has been added for the found of the focusing matrix. RSS method is to make sure that the error between array manifold after focusing and that at the reference frequency point is minimum in the condition that focusing matrix is unitary, that is

$$\begin{cases} \min\limits_{\boldsymbol{T}(f_j)} \left\| \boldsymbol{A}(f_0) - \boldsymbol{T}(f_j)\boldsymbol{A}(f_j) \right\|_F^2, \quad j = 1,\ 2, \cdots, J \\ \text{s.t.} \qquad \boldsymbol{T}(f_j)\boldsymbol{T}^{\mathrm{H}}(f_j) = \boldsymbol{I} \end{cases} \tag{5}$$

where $\|\cdot\|_F$ is the Frobenius model, in fact, the optimal problem above is to solve the least square solution in the condition of unitary matrix, so solve the equation above, we have

$$\boldsymbol{T}(f_j) = \boldsymbol{V}_0 \boldsymbol{V}_j^{\mathrm{H}} \tag{6}$$

where $\boldsymbol{V}_0$ and $\boldsymbol{V}_1$ are respectively the matrix composed by the left and right singular vectors of $\boldsymbol{A}(f_0)\boldsymbol{A}^{\mathrm{H}}(f_j)$ after singular value decomposition (SVD).

Another problem is choosing the reference frequency f_0, we can choose the center frequency of the signal as the reference point in many cases, but maybe it is not optimal, one of the best selection can be solved as follows:

$$\begin{cases} \min\limits_{f_0} \sum\limits_{j=1}^{J} w_j \min\limits_{\boldsymbol{T}(f_j)} \left\| \boldsymbol{A}(f_0) - \boldsymbol{T}(f_j)\boldsymbol{A}(f_j) \right\|_F^2 \\ \text{s.t.} \qquad \boldsymbol{T}^H(f_j)\boldsymbol{T}(f_j) = \boldsymbol{I} \qquad j = 1, 2, \cdots, J \end{cases} \tag{7}$$

where w_j is a normalized weighting coefficient who is proportional to the SNR of the *j*th frequency component, when the power of the signal in the band is flat, define $w_j = \dfrac{1}{J}$. The equation (7) can be changed into the solution of the problem below

$$\max_{f_0} \sum_{i=1}^{K} \mu_i \sigma_i \left(\boldsymbol{A}(f_0) \right) \tag{8}$$

where $\mu_i = \sum\limits_{j=1}^{J} w_j \sigma_i \left(\boldsymbol{A}(f_j) \right)$, and $\sigma_i(\cdot)$, $i = 1, 2, \cdots, K$ means singular values of matrix who arranges in non increasing order, K is the number of column of direction matrix. The equation above is one dimensional optimizing problem, the optimal reference frequency f_0 can be easily acquired by searching frequency space.

3.2 Proposed Method

According to the model of the wideband signal, the array covariance matrix of frequency f_i can be expressed as

$$\begin{aligned} \boldsymbol{R}_{XX}(f_i) &= E\left[\boldsymbol{X}(f_i)\boldsymbol{X}^{\mathrm{H}}(f_i) \right] \\ &= \boldsymbol{A}(f_i)\boldsymbol{R}_{SS}(f_i)\boldsymbol{A}^{\mathrm{H}}(f_i) + \sigma^2 \boldsymbol{I} \end{aligned} \tag{9}$$

where

$$\boldsymbol{R}_{SS}(f_i) = E\left[\boldsymbol{S}(f_i)\boldsymbol{S}^{\mathrm{H}}(f_i) \right] \tag{10}$$

here, a theorem is listed below [8]:

Theorem $\boldsymbol{A}, \boldsymbol{B} \in C^{M \times L}$ $(L \le M)$ are two arbitrary matrices with full column rank, they satisfy the following minimization problem:

$$\boldsymbol{T}_{\text{opt}} = \arg\left\{ \min_{\boldsymbol{T}} \left\| \boldsymbol{TA} - \boldsymbol{B} \right\|_F^2 \quad \text{s.t. } \boldsymbol{T}^{\mathrm{H}}\boldsymbol{T} = \boldsymbol{I}_M \right\} \tag{11}$$

where $T \in C^{M\times M}$, when $L > M-1$, there is only one solution for the equation above; when $L = M-1$, there are two solutions; when $L < M-1$, there are infinite solutions. Specifically, define

$$\boldsymbol{C} \triangleq \boldsymbol{A}\boldsymbol{B}^{\mathrm{H}} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^{\mathrm{H}} = [\boldsymbol{U}_1 \ \ \boldsymbol{U}_2] \begin{bmatrix} \boldsymbol{\Lambda}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{V}_1^{\mathrm{H}} \\ \boldsymbol{V}_2^{\mathrm{H}} \end{bmatrix} \tag{12}$$

it is the singular value decomposition (SVD) of the matrix $\boldsymbol{C}$, then the disaggregation can be expressed as

$$\boldsymbol{T}_{\mathrm{opt}} = \boldsymbol{V}\boldsymbol{U}^{\mathrm{H}} = \boldsymbol{V}_1\boldsymbol{U}_1^{\mathrm{H}} + \boldsymbol{V}_2\boldsymbol{U}_2^{\mathrm{H}} \tag{13}$$

where $\boldsymbol{U}_1, \boldsymbol{V}_1 \in C^{M\times L}$ is the only certain unitary matrix, and $\boldsymbol{U}_2, \boldsymbol{V}_2 \in C^{M\times(M-L)}$ is the arbitrary matrix who satisfy the property below:

$$\begin{cases} \boldsymbol{U}_2^{\mathrm{H}}\boldsymbol{U}_2 = \boldsymbol{I}_{M-L} \\ \boldsymbol{U}_2^{\mathrm{H}}\boldsymbol{U}_1 = \boldsymbol{0}_{M-L,\,L} \\ \boldsymbol{V}_2^{\mathrm{H}}\boldsymbol{V}_2 = \boldsymbol{I}_{M-L} \\ \boldsymbol{V}_2^{\mathrm{H}}\boldsymbol{V}_1 = \boldsymbol{I}_{M-L,\,L} \end{cases} \tag{14}$$

where $\boldsymbol{0}_{M\times(M-L)}$ is $M\times(M-L)$ dimensional matrix.

It can be deduced from the theorem above, RSS algorithm is a special situation of the optimal problem of equation (11), if the number of signal is smaller than that of the array, the solution is not single, so a further constraint can be added to ensure the uniqueness of the focusing matrix and a better performance. This idea will be used for design of the focusing matrix based on R-CSM, so the constrain can be described as: in the view of some particular partition the total focusing effect is the best, that is to increase the robustness of the focusing matrix to incident angle.

The R-CSM is made up of two classes of focusing matrix, they respectively use for the different stages of iterative estimation. The first one is used for the initial stage, there is no any priori knowledge about the DOA, so it must be designed for the matrix which is still effect under the circumstance of no any priori knowledge; The second one is used from the second stage, it has the DOA information obtaining from the last stage, if we want to acquire the better performance to calculate the true DOA, it is necessary to enhance the robustness of the focusing matrix of different direction variable.

The first kind of focusing matrix at the initial stage is obtained by solving the optimal problem below:

$$\begin{cases} \boldsymbol{T}_j[0] = \arg\left\{ \min\limits_{\boldsymbol{T}} \int_{-\frac{1}{2}}^{\frac{1}{2}} w(u) \left\| \boldsymbol{T}\boldsymbol{a}(u, f_j) - \boldsymbol{a}(u, f_0) \right\|_F^2 \mathrm{d}u \right\} \\ \qquad \text{s.t.} \qquad \boldsymbol{T}^{\mathrm{H}}\boldsymbol{T} = \boldsymbol{I}, \quad j = 1,\ 2, \cdots, J \end{cases} \tag{15}$$

where $\boldsymbol{T} \in C^{N\times N}$, N is the number of sensors, $\boldsymbol{a}(u, f_j)$ is the array manifold at frequency f_j and spatial frequency $u = \dfrac{x\cos\theta\cos\phi + y\sin\theta\cos\phi}{\lambda}$, f_0 is the focusing frequency, $w(u)$ is the normalized weighted function, here define $w(u) = 1$. Its physical significance is to minimize the focusing error in the whole angle space, the optimal solution of the problem above is given from reference [8]

$$\boldsymbol{T}_j[0] = \boldsymbol{V}_j \boldsymbol{U}_j^{\mathrm{H}} \tag{16}$$

where $\boldsymbol{V}_j$ and $\boldsymbol{U}_j$ can be acquired from SVD of the matrix below

$$\boldsymbol{Q}_j = \int_{-\frac{1}{2}}^{\frac{1}{2}} \boldsymbol{a}(u, f_j)\boldsymbol{a}^{\mathrm{H}}(u, f_0) w(u)\, \mathrm{d}u = \boldsymbol{U}_j \boldsymbol{\Lambda}_j \boldsymbol{V}_j^{\mathrm{H}} \tag{17}$$

After focusing matrix $\boldsymbol{T}_j[0]$ is used at the initial stage, we can obtain a group of initial estimation of DOA. After the second iteration, the second kind of focusing matrix is used, it can be obtained by solving the following two optimal problems, a kind of matrix can be calculated according to the theorem

$$\begin{cases} \boldsymbol{T}_j[i] = \arg\left\{ \min\limits_{\boldsymbol{T}} \left\| \boldsymbol{T}\boldsymbol{A}\left(\hat{\boldsymbol{u}}[i-1], f_j\right) - \boldsymbol{A}\left(\hat{\boldsymbol{u}}[i-1]\right), f_0 \right\|_F^2 \right\} \\ \quad \text{s.t.} \boldsymbol{T}^{\mathrm{H}}\boldsymbol{T} = \boldsymbol{I}, \quad j = 1,\ 2, \cdots, J \end{cases} \tag{18}$$

where $\hat{\boldsymbol{u}}\,[i-1]$ is spatial frequency vector of the (i-1)th step. Then the only robust focusing matrix is estimated by solving the following problem:

$$\begin{cases} \boldsymbol{T}_j[i] = \arg\left\{ \min\limits_{\boldsymbol{T}_j[i]} \sum\limits_{k=1}^{K} \int\limits_{v_k[i]} \left\| \boldsymbol{T}_j[i]\, \boldsymbol{a}\left(u, f_j\right) - \boldsymbol{a}(u, f_0) \right\|_F^2 w(u)\, \mathrm{d}u \right\} \\ \quad \text{s.t.} \quad \boldsymbol{T}_j^{\mathrm{H}}[i]\boldsymbol{T}_j[i] = \boldsymbol{I}, \quad j = 1,\ 2, \cdots, J \end{cases} \tag{19}$$

where $v_k[i]$ is called robust region. After the first step of iteration, there is no need to focus the array manifold of the whole view area, the more degree of freedom can be used for the more robustness to the estimated spatial frequency. So the robust region should be diminishing as the increase of the iteration times, and the spatial frequency of the last estimation is defined as the region center, one of the robust area can be selected as:

$$v_k[i] = \left[\max\left\{ -\frac{1}{2}, \hat{u}_k[i-1] - \frac{1}{2i^p} \right\}, \min\left\{ \frac{1}{2}, \hat{u}_k[i-1] + \frac{1}{2i^p} \right\} \right] \tag{20}$$

where $\hat{u}_k[i-1]$ is the spatial frequency of the kth signal acquired from the last iteration, p>0 is a variable related to the convergence speed of the robust area, and it can be optimized, here define p=2.

The solution of equation (19) is

$$\boldsymbol{T}_j[i] = \boldsymbol{V}_1 \boldsymbol{U}_1^{\mathrm{H}} + \boldsymbol{Y}\boldsymbol{F}\boldsymbol{G}^{\mathrm{H}}\boldsymbol{X}^{\mathrm{H}} \tag{21}$$

where $\boldsymbol{V}_1$ and $\boldsymbol{U}_1$ are obtained from the following SVD

$$\boldsymbol{A}(\hat{\boldsymbol{u}}, f_j)\boldsymbol{A}^{\mathrm{H}}(\hat{\boldsymbol{u}}, f_0) = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^{\mathrm{H}} = \begin{bmatrix} \boldsymbol{U}_1 & \boldsymbol{U}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\Lambda}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{V}_1^{\mathrm{H}} \\ \boldsymbol{V}_2^{\mathrm{H}} \end{bmatrix} \tag{22}$$

where all the columns of $\boldsymbol{X}$ form a group of orthogonal basis in orthogonal complementary space of $\boldsymbol{U}_1$, that of $\boldsymbol{Y}$ form a group of orthogonal basis in orthogonal complementary space of $\boldsymbol{V}_1$, the matrix $\boldsymbol{F}$ and $\boldsymbol{G}$ are obtained from the SVD below:

$$\boldsymbol{Y}^{\mathrm{H}}\boldsymbol{Q}^{\mathrm{H}}\boldsymbol{X} = \boldsymbol{F}\boldsymbol{\Delta}\boldsymbol{G}^{\mathrm{H}} \tag{23}$$

where

$$\boldsymbol{Q} = \sum_{k}^{K} \int_{v_k} \boldsymbol{a}(u, f_j)\boldsymbol{a}^{\mathrm{H}}(u, f_0)w(u)\mathrm{d}u \tag{24}$$

Thus, the method is summarized as follows:

Step1: Apply a DFT to the array output to sample the spectrum of data, then use equation (9) to solve the covariance matrix of every frequency, then determine the reference frequency f_0;

Step2: Use SVD to the covariance matrix of every frequency, then calculate focusing matrix $\boldsymbol{T}_j[0]$ according equation (21);

Step3: Focus the covariance matrices of every frequency on the reference point f_0, then estimate initial DOA by the method adaptive to narrowband signal;

Step4: Take the initial estimating result into equation (20) to compute robust area;

Step5: Calculate the further focusing matrices by equation (21), (22), (23), (24);

Step6: Repeat the step 3 to 5 until the result converge.

As the method adapts to two-dimensional wideband signals, so we can call it TDR-CSM method for short.

4 Simulations

In order to verify the effective of the APA method, three simulations are presented with matlab below, consider some wideband chirp signals impinge on 8 arbitrary placed plane array, their coordinates are (0, 0), (-0.15, 0.15), (-0.078, 0.22), (-0.2, 0.061), (-0.23, -0.049), (0.053, 0.12), (0.22, 0.08), (0.066, -0.041), it is in meters, the center frequency of the signals is 3GHz, the width of the band is 20% of the center frequency, TDR-CSM and RSS methods are respectively used for the simulations, and comparing with their spatial spectrum figures, resolution probability and angle measurement accuracy, the center frequency of the signals is selected as the reference frequency.

In the first simulation, two wideband coherent signals with the same power impinge on the array from directions $(30^\circ, 40^\circ)$ and $(50^\circ, 60^\circ)$, the snapshots of

every frequency is 100, 300 times Monte-Carlo simulations have run for each iteration, take the average value as the final result. Fig.2 and Fig.3 respectively show the Root mean square error (RMSE) of the estimation when SNR are 3dB and 8dB below.

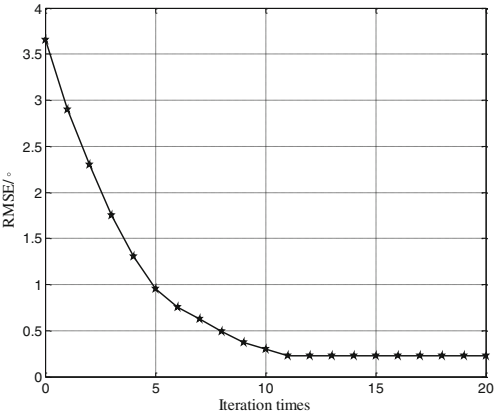


Fig. 2 RMSE versus iteration times when SNR is 3dB

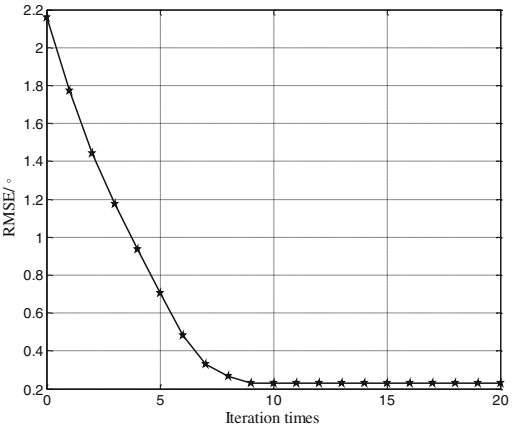


Fig. 3 RMSE versus iteration times when SNR is 8dB

It is seen from the Fig.2 and Fig.3, when SNR is 3dB, RMSE is tending towards stability as iteration times is 11; and when SNR is 8dB, RMSE is tending towards stability as iteration times is 9, that is to say the iteration times under circumstance of high SNR is less than that of low SNR, we can modify the iteration times according to the actual SNR.

In the second simulation, two wideband coherent signals with the same power impinge on the array from directions $(30^\circ, 40^\circ)$ and $(50^\circ, 60^\circ)$, the snapshots of every frequency is 100, 300 times Monte-Carlo simulations have run for each iteration, take the average value as the final result. Fig.4 shows the Root mean

square error (RMSE) versus SNR with the methods of TDR-CSM and RSS below, where iteration times of TDR-CSM is 12.

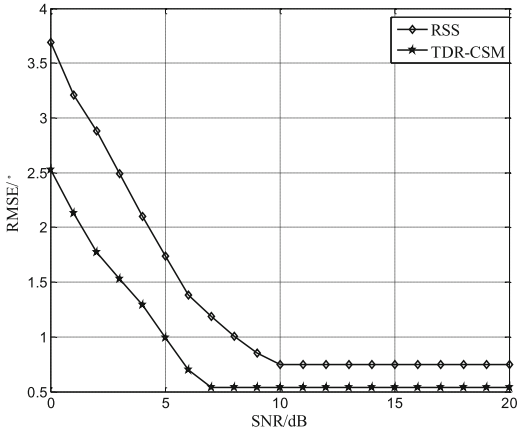


Fig. 4 RMSE of different methods versus SNR

It is seen from Fig.4, the RMSE of the two methods is tending towards stability as SNR increases, and the performance of TDR-CSM is better than RSS.

5 Conclusions

The paper proposed a new method of founding focusing matrix for two-dimensional wideband signals, it overcomes the shortcoming of need for pre-estimate to the direction, the focusing matrix is built by the process of iteration, it needs to form a group of robust focusing matrix to the DOA in every step. We can see from the simulation, as the robust area decreases of the TDR-CSM along with the iteration, its performance is improving, besides, the method has low demand for the position of the plane array, it adapts to arbitrary plane array for two-dimensional DOA estimation.

Acknowledgments. This work is supported by National Natural Science Foundation of China (no. 61201399), China Postdoctoral Science Foundation (no. 2012M511003), Project of Science and Technology of Heilongjiang Provincial Education Department (no.12541638), Youth Foundation of Heilongjiang University (no.201026), and Startup Fund for Doctor of Heilongjiang University.

References

1. Schmidt, RO: Multiple emitter location and signal parameter estimation. In: Conference of RADC Spectrum Estimation Workshop Control, vol. 34(3), pp. 276–280 (1979)

2. Roy, K.T.: Esprit-estimation of signal parameters via rotational invariance techniques. IEEE Transactions on Acoustic, Speech and Signal Processing **37**(7), 984–995 (1989)
3. Su, G., Morf, M.: Signal subspace approach for multiple wideband emitter location. IEEE Transactions on Acoustics, Speech and Signal Processing **31**(6), 1502–1522 (1983)
4. Salman, N., Ghogho, M., Andrew, H.: On the Joint Estimation of the RSS-Based Location and Path-loss Exponent. IEEE Wireless Communications Letters **1**(1), 34–37 (2012)
5. Hung, H., Kaveh, M.: Focussing matrices for coherent signal-subspace processing. IEEE Transactions on Acoustics, Speech, and Signal Processing **36**(8), 1272–1281 (1988)
6. Hong, W., Tewfik, A.H.: Focusing matrices for wideband array processing with no a priori angle estimates. IEEE Transactions on Acoustic, Speech and Signal Processing **2**, 1292–1302 (1992)
7. Valaee, S., Kabal, P.: Wideband array processing using a two-sided correlation transformation. IEEE Transactions on Signal Processing **43**(1), 160–172 (1995)
8. Sellone, F.: Robust auto-focusing wideband DOA estimation. Signal Processing **86**(1), 17–37 (2006)
9. Feng, D., Bao, M., Ye, Z., Guan, L., Li, X.: A novel wideband DOA estimator based on Khatri-Rao subspace approach. Signal Processing **91**(10), 2415–2419 (2011)
10. Huang, K., Yiyu, Z., Guozhu, Z.: A fast DOA estimation method of wideband signals based on Krylov subspace. Journal of Astronautics **26**, 461–465 (2005)
11. Palanisamy, P., Kalyanasundaram, N., Raghunandan, A.: A new DOA estimation algorithm for wideband signals in the presence of unknown spatially correlated noise. Signal Processing **89**(10), 1921–1931 (2009)
12. Jin, Z., Zhongfu, Y., Yunxiang, M.: An efficient DOA estimation method for wideband coherent signals in the presence of correlated noise. Acta Electronica Sinica **41**, 1278–1283 (2013)

Part II
QoS Control and Assessment in Networked Multimedia Applications

Network Adaptive Flow Control Algorithm for Haptic Data Over the Internet–NAFCAH

George Kokkonis, Kostas E. Psannis and Manos Roumeliotis

Abstract This paper deals with the transfer of real time haptic data over the Internet. Some interested transport protocols have already been proposed for the transport of real time haptic data. This paper presents the related work on haptic data transferring. A new network adaptive flow control algorithm is proposed. The new algorithm combines most of the known flow control algorithms while taking into account the network conditions of the Internet and the significant haptic events.

Keywords Haptics · Tele-Haptics · Transport protocols · Teleoperation · Interative applications · Real time protocol · Flow control · Congestion control

1 Introduction

Real time data were considered until recently only video and audio data. With the optimization of telerobotics and the improvement of Internet status, a new kind of data made its appearance the last decade. This is tele-haptic data. With the word haptics we refer to the tactile and kinesthetic human sense. As human population grows older, the need for teleoperation is getting bigger. With the help of tele-haptics some risky jobs, such as nuclear disposal and wreckage exploration could be made with great safety. Furthermore, applications as tele-surgery, tele-mentoring, haptic video games, and augmented reality are only few examples of the many sectors of our daily life that tele- haptics could be applied to.

The main obstacle that impedes tele-haptics from flourishing is the delay and the jitter that is being encountered in the Internet. Several congestion/flow control algorithms have been proposed for the limitation of the negative effects of the delay and jitter. Some of them are the TCP congestion window [1], the Additive Increase/Multiplicative Decrease AIMD[2], the Rate Based Congestion Control RAP [3], and the TCP Friendly Rate Control (TFRC) [4], and the variable Inter packet Gap (IPG) [5].

G. Kokkonis(✉) · K.E. Psannis · M. Roumeliotis
Department of Applied Informatics, University of Macedonia,
156 Egnatia Street, 54006 Thessaloniki, Greece
e-mail: {gkokkonis,kpsannis,manos}@uom.gr

T.T. Zin et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 388,
DOI: 10.1007/978-3-319-23207-2_10

Apart from the common congestion control algorithms some rather interesting technics intent to reduce the transfer rate of the haptic stream such as the packetization intervals [6], the differential coding with quantization [7], the haptic event prioritization [8], the dead-reckoning [9] and the perception based compression using Kalman filters [10].

Furthermore, the adaptive buffering [9], the haptic packet prioritization [11] and the wave variables [12] try to mitigate the jitter and the delay of the network.

The rest of the paper is organized as follows. Section 2 presents the related work on congestion/flow control algorithms that could be applied to haptic applications and describes network conditions that should be fulfilled for a satisfying QoE. Section 3 analyzes the new proposed flow control algorithm for tele-haptic applications. Section 4 presents graphical representations of experimental data. Finally section 5 identifies conclusions and future work.

2 Related Work on Congestion/Flow Control Algorithms

Lot of research has been done for the mitigation of the negative effects of the network delay and jitter. One way to avoid congestion is by minimizing the transmitted haptic packets when delay and jitter show increasing signs [5]. Another method tries to minimize the transmitted packets by forcing the receiver to predict the packets that hasn't receive [10]. A further technique transmits only the packets that produce haptic feedback perceptible form human senses [9]. Other methods try to compress the haptic data with lossy data reduction techniques such as the quantization and the differential coding [13]. Furthermore, some researchers send the haptic packets with different priorities [8], the most important packets are sent with higher priority and more reliable than other packets. One more method for minimizing the transfer rate is the packetization intervals [6], where a number of packets are grouped together in a frame and sent to the receiver as a packet.

Each of the above techniques presents some advantages regarding bandwidth, packet loss, and jitter at the expense of precision and average delay. Depending on the application most of the above techniques improve Quality of Experience (QoE) of the user at specific network conditions.

Table 1 QoS Requirements for Multimedia Streams [14 - 18]

	APPLICATIONS			
QOS	***HAPTICS***	***VIDEO***	***AUDIO***	***GRAPHICS***
JITTER (ms)	≤ 2	≤ 30	≤ 30	≤ 30
DELAY (ms)	≤ 50	≤ 400	≤ 150	≤ 100-300
PACKET LOSS (%)	≤ 10	≤ 1	≤ 1	≤ 10
UPDATE RATE (Hz)	≥ 1000	≥ 30	≥ 50	≥ 30
PACKET SIZE (bytes)	64-128	≤ MTU	160-320	192-5000
THROUGHPUT (Kbps)	512-1024	2500-40000	64-128	45-1200

The preferred network conditions may vary from application to application. Many studies [14 - 18] conclude that the network condition should satisfy the limitation of Table 1, in order the QoE of the user to be satisfactory.

3 The Proposed Flow Control Algorithm-NAFCAH

Since the network conditions of the Internet are time-varying, the algorithm that controls the transmission of the packets should be network adaptive. Apart from the adaptive transmission rate and bandwidth, priority should be enforced in the haptic packets. Some packets are more important than others. These packets should be sent with higher priority and more reliably. If the network conditions are deteriorating some packets with lower priority should not be sent at all. Another metric that should be network adaptive is the size of the transmitted packets. Techniques as the differential coding and the quantization, should modify their parameters, in order to change the packet size, and as a consequence the bandwidth of the haptic stream. The transmission rate of the packets should be network adaptive as well. If the network shows some little signs of congestion as increased delay, jitter and packet loss, then the transmission rate should be reduced in order heavy congestion to be avoided.

Apart from the maximum values of the network delay, jitter and packet loss of Table 1, intermediate values should be established, in order to escalate the QoS and as a consequence the QoE of the users. The maximum values of Table I should be escalated into three values. The first scale of these values is [0-max/3] that corresponds to perfect network conditions and the QoE should increase in order to reach its maximum value. The scale (max/3 –2*max/3] corresponds to fair network conditions where the QoE should increase slowly. The scale (2*max/3 –max] corresponds to acceptable conditions but with high possibility of diversion. The flow algorithm should try to avoid this diversion by keeping the QoE steady. When network conditions are worse than the maximum allowable value of Table 1, the flow algorithm should lower the QoE rapidly, so as to avoid congestion. One way to measure the network conditions of the Internet is by sending periodically ICMP packets over the UDP protocol from the sender to the receiver. The delay of the ICMP packets corresponds to the delay of the network d_{net}.

Base on the above assumptions the Network Adaptive Flow Control Algorithm for Haptic data – NAFCAH is proposed. The system model of the NAFCAH is depicted in Fig. 1.

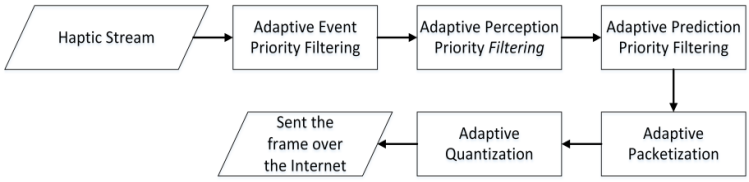


Fig. 1 System model of Network Adaptive Flow Control Protocol –NAFCAH

3.1 Network Adaptive Packet Priority

One way to reduce the transmission rate and bandwidth of the haptic stream is to reduce the packets that are being transferred. Packets with lower priority can sometimes be omitted in order to reduce the transfer rate and bandwidth. In [8] the haptic event priority is introduced. When the Haptic Interaction Pointer (HIP) of the haptic interface approaches a virtual object, the haptic packet should obtain higher Event Priority *ep*.

Event Priority ep

When the distance *dis* between the HIP and the virtual object decreases, the event priority *ep* increases based on the equation (1) and depicted in Fig. 2. The event priority takes its maximum value *ep=1* when the HIP touches the virtual object.

$$ep = \begin{cases} 1 & ,dnet < dmax/3 \\ 1/(n * dis + 1) & ,dmax/3 < dnet < 2dmax/3 \\ 1/(2 * n * dis + 1) & ,2dmax/3 < dnet < dmax \\ 1/(3 * n * dis + 1) & ,dmax < dnet \end{cases} \quad (1)$$

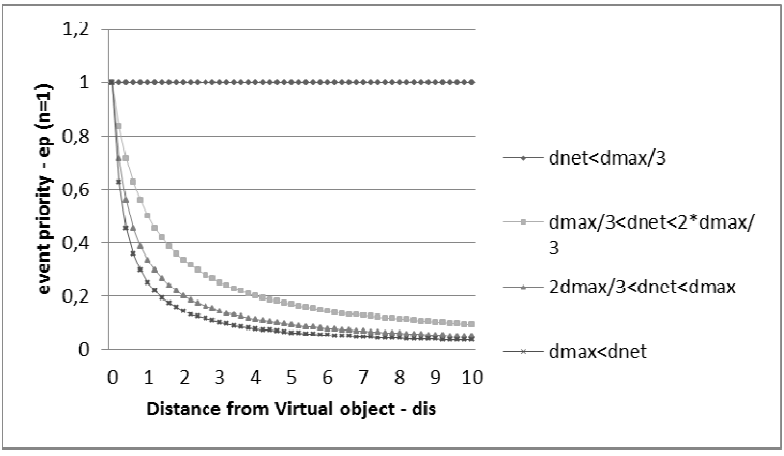


Fig. 2 Event priority vs Distance from Virtual object

The higher the priority *ep* is, the higher the importance of the packet to be transmitted. When the priority *ep* is equal to 1, all the corresponding packets should pass the priority check and should be forwarded to the next priority process. The factor n>0, is set by the user and represents how steep curve of Fig 2 and 3 will be. When the factor *ep* is smaller to 1, the HIP doesn't encounter any impedance and the user relies on his/her visual sense to handle the haptic interface. The update rate in this case could be much lower than 1 KHz, equal to the necessary update rate for the vision sense 30 Hz. If network conditions are not adequate, intermediate packets should be dropped. Based on this observation, the

update rate of packets that pass the event priority check *pr* is described in equation (2) and depicted in Fig. 3.

$$pr = 970 * ep + 30 \tag{2}$$

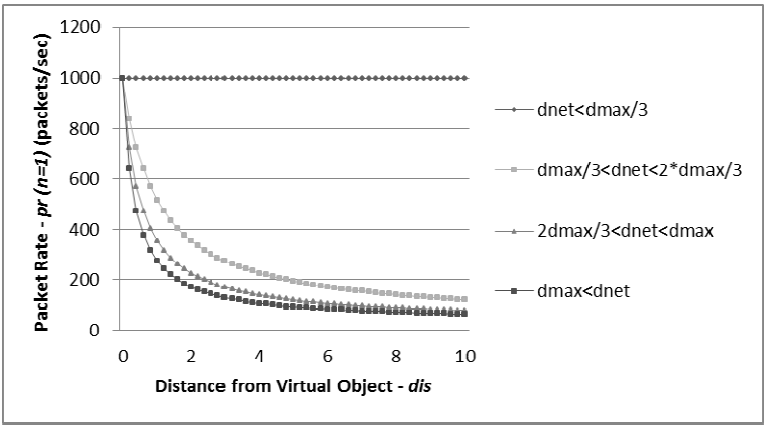


Fig. 3 Packet Rate vs Distance from Virtual object

Perception Priority *pep*

Packets that don't generate perceptible haptic feelings to the user [9] should have lower Perceptible Priority *pep*. The perceptible priority *pep* is based on the dead-reckoning theory[9]. It uses the Weber's law of Just Noticeable Difference (JND) [19] to calculate the threshold ΔI. Equation (3) calculates the JND based on the stimulus intense I, that the haptic interface causes to the user. The constant κ is called the Weber fraction.

$$\Delta I = I * \kappa \tag{3}$$

When the haptic packet produces difference on the stimulus intense dI higher than ΔI then the packet should transmitted with high priority. If the dI is lower the ΔI the packet should have lower priority on the transmission.

In order the algorithm to be network adaptive, the Weber fraction κ_i should change according to the network conditions based on equation (4). When the network conditions deteriorate, the constant κ_i should increase in order fewer, but more important, packets to be transmitted. The factor $0<h<1$ is set by the user and represents how rapid the alteration of κ will be. The factor κ_i will try to change its value every time there is a feedback from the network for its network conditions. As far as the Weber fraction κ_0 is concerned Karadogan et al. in [20] have shown that the Weber fraction should take values from $0.08<\kappa_i<0.3$ depending on the application. The smaller the Weber fraction is, the higher the QoE of the user.

$$\kappa_i = \begin{cases} \kappa_0 = 0.08 & \\ \kappa_{i-1} * (1 - 2 * h) & , dnet < dmax/3 \\ \kappa_{i-1} * (1 - h) & , dmax/3 < dnet < 2dmax/3 \\ \kappa_{i-1} & , 2dmax/3 < dnet < dmax \\ \kappa_{i-1} * (1 + 2 * h) < 0.3 & , dmax < dnet \end{cases} \quad (4)$$

The packets that pass the perception priority filtering, are passing to the next stage of the sorting.

Prediction Priority *pp*

Packets that can be predicted by previous packets [7] should obtain a lower Prediction Priority *pp*. Several Interesting studies [10] have shown that most of the transmitted packets could be predicted based on the previous data. If the movement of the HIP is linear the prediction is precise. The prediction unit is installed both at the sender and at the receiver. If the prediction unit at the sender calculates that the current packet could be predicted at the receiver from the last packets that were send, then the current packet is not transmitted. In order the algorithm to be network adaptive, apart from the identical predicted packets, predicted packets that are similar to the real packets could be excluded from the transmission. Again this algorithm should be based on Weber's law of the JND to decide which packets could successfully be predicted at the receiver side. If the predicted packet doesn't produce greater difference on the stimulus intense dI to the users from the real packet than the threshold $\Delta I'$, then the packet is not transmitted, but it will be predicted on the receiver's side. In this case, the Weber fraction m_i could have different value from the variableκ of equation (3). Again the variable m_i should be changed according to the network conditions. The equations that decide which packets are predicted successfully are depicted in (5) and (6). The factor $0<j<1$ is set by the user and represents how rapid the alteration of m will be.

$$\Delta I_i' = I * m_i \quad (5)$$

$$m_i = \begin{cases} m_0 = 0.08 & \\ m_{i-1} * (1 - 2 * j) & , dnet < dmax/3 \\ m_{i-1} * (1 - j) & , dmax/3 < dnet < 2dmax/3 \\ m_{i-1} & , 2dmax/3 < dnet < dmax \\ m_{i-1} * (1 + 2 * j) < 0.3 & , dmax < dnet \end{cases} \quad (6)$$

3.2 Adaptive Transmission Rate

Wirz et al. have shown that the network adaptive transmission rate of the haptic stream improves teleoperation [5]. The main obstacle in the variation of the transmission rate is the stable production of packages in the source. If the haptic interface produces update packets steadily and the sender fluctuates the sending rate, a buffer is necessary at the sender side to absorb the fluctuation [9].

The negative aspect of this technique is that if the haptic interface produces packets at very high update rate *ur*, usually 1 KHz [21], the sender should transmit its packets sometimes even faster to compensates the previous lower rates. This even higher update rate often results in congestion and packet loss. In order to lower the update rate, an interesting proposal is to integrate a group of packets in a frame and sent them as a unified packet, a technic called packetization interval. Fujimoto and Ishibashi [6] have proven that a packetization interval of np_{max}=8 packets that is sent every 8 ms improves the systems performance in overloaded networks. Another interesting study [11] has shown that the number of the integrated packets np_i should vary, depending on the network delay, in order not to overcome the maximum allowable delay d_{max}. Every packet that is integrated in the frame adds 1/*ur* sec of delay. If we take into account the network delay d_{net}, then the maximum number of integrated packets $np_{i,max}$ is:

$$np_{i,max} = (d_{max} - d_{net}) * ur \tag{7}$$

The number of integrated packets is described at equation (8).

$$np_i = \begin{cases} np_0 = 8 & \\ np_{i-1} - 2 > 0 & , dnet < dmax/3 \\ np_{i-1} - 1 > 0 & , dmax/3 < dnet < 2 * dmax/3 \\ np_{i-1} & , 2dmax/3 < dnet < dmax \\ np_{i-1} + 2 < 8 & , dmax < dnet \end{cases} \tag{8}$$

3.3 Network Adaptive Quantization

The bandwidth that a haptic stream absorbs, depends on two factors, the frame rate, that is determined from equation (8) and the size of the frame. The frame size could be reduced if differential coding and quantization [13] is enforced on the packets that are grouped in the frame. The technique that is recommended is the Differential Pulse-Code Modulation (DPCM). In case of a slow motion of the HIP, most of the haptic packets have similar values. The differential coding will produce smaller values than the original ones. Smaller values mean fewer bits. The quantization of the differentiate values could be made with variable quantization step qs_i. The Adaptive Differential Pulse-Code Modulation (ADPCM) for haptic packets was introduced in [13]. Cyrus et al. proposed to alter the quantization step according to the difference size. In this paper the authors propose to change the quantization step according to the network conditions. When the network conditions deteriorate the qs_i should increase in order fewer bits to be required for the reconstruction of the original values. The qs_i is calculated based on equation (9). The factor $0<l<1$ indicates how rapid the alteration of qs_i will be, in accordance to the network feedback.

$$qs_i = \begin{cases} qs_0 = 0.3 \\ qs_{i-1} * (1 - 2 * l) & , dnet < dmax/3 \\ qs_{i-1} * (1 - l) & , dmax/3 < dnet < 2dmax/3 \\ qs_{i-1} & , 2dmax/3 < dnet < dmax \\ qs_{i-1} * (1 + 2 * l) < 1 & , dmax < dnet \end{cases} \quad (9)$$

The initial quantization step qs_0 in [22] after experiment regarding Mean Opinion Score (MOS) is recommended $qs_0=0.3\ mm$.

The flowchart of Network Adaptive Flow Control Protocol -NAFCAH which is based on the above priorities and compressions is depicted in Fig. 4.

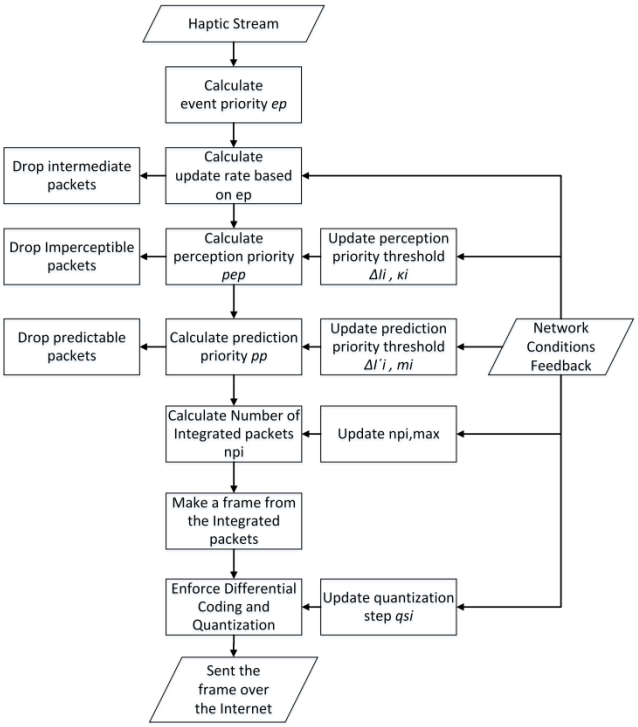


Fig. 4 Flowchart of Network Adaptive Flow Control Protocol –NAFCAH

4 Conclusions and Future Work

It is known that the network conditions of the Internet are constantly changing. The metrics such as the network delay, the jitter and the packet loss are not stable. In such a variable environment a flow control algorithm for transferring haptic data is necessary. In this paper a network adaptive flow control algorithm named NAFCAH is presented. It is a quit flexible algorithm where the user can adjust its sensitivity to the network variations by configuring the factors *n, h, j, l*. All the

known congestion and flow control techniques have been enforced, in order to achieve the desired result. Packet priorities such as the event priority, the perception priority and the prediction priority are described and defined. Packetization Interval technique and lossy compression methods such as the Adaptive Differential Pulse-Code Modulation are enforced.

It has already been scheduled to evaluate the presented flow control algorithm in simulations and real world experiments. These experiments will have as a primary target to define the proposed values for the factors *n, h, j, l,* and the initial and the maximum values of κ, *m,qs* that the user should set for the sensitivity of the flow control algorithm.

References

1. Gerla, M., Sanadidi, M.Y., Wang, R., Zanella, A., Casetti, C., Mascolo, S.: Tcpwestwood: congestion window control using bandwidth estimation. In: IEEE Global Telecommunications Conference, GLOBECOM 2001, vol. 3, pp. 1698–1702. IEEE (2001)
2. Yang, Y.R., Lam, S.S.: General aimd congestion control. In: Intern. Conf. on Network Protocols, pp. 187–198. IEEE (2000)
3. Rejaie, R., Handley, M., Estrin, D.: Rap: an end-to-end rate-based congestion control mechanism for real-time streams in the internet. In: Eighteenth Annual Joint Conf. of the IEEE Computer and Communications Societies, INFOCOM 1999, vol. 3, pp. 1337–1345. IEEE (1999)
4. Widmer, J., Denda, R., Mauve, M.: A survey on tcp-friendly congestion control. IEEE Network **15**(3), 28–37 (2001)
5. Wirz, R., Marn, R., Ferre, M., Barrio, J., Claver, J.M., Ortego, J.: Bidirectional transport protocol for teleoperated robots. ÍEEE Trans. on Industrial Electronics **56**(9), 3772–3781 (2009)
6. Fujimoto, M., Ishibashi, Y.: Packetization interval of haptic media in networked virtual environments. In: Proc. of 4th ACM SIGCOMM Workshop on Network and System Support for Games, pp. 1–6. ACM (2005)
7. Borst, C.W.: Predictive coding for efficient host-device communication in a pneumatic force-feedback display. In: Eurohaptics Conf., Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems, pp. 596–599. IEEE (2005)
8. Lee, S., Kim, J.: Priority-based haptic event filtering for transmission and error control in networked virtual environments. Multimedia systems **15**(6), 355–367 (2009)
9. Ishibashi, Y., Hashimoto, Y., Ikedo, T., Sugawara, S.: Adaptive delta-causality control with adaptive dead-reckoning in networked games. In: Proc. of the 6th ACM SIGCOMM Workshop on Network and System Support for Games, pp. 75–80. ACM (2007)
10. Hinterseer, P., Steinbach, E., Chaudhuri, S.: Perception-based compression of haptic data streams using kalman filters. In: IEEE Intern. Conf. on Acoustics, Speech and Signal Processing, vol. 5, pp. V–V. IEEE (2006)
11. Lee, S., Kim, J.: Dynamic network adaptation scheme employing haptic event priority for collaborative virtual environments. In: Proc. of the First International Conference on Immersive Telecommunications. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), p. 12 (2007)

12. Aziminejad, A., Tavakoli, M., Patel, R.V., Moallem, M.: Transparent time-delayed bilateral teleoperation using wave variables. IEEE Trans. on Control Systems Technology **16**(3), 548–555 (2008)
13. Shahabi, C., Ortega, A., Kolahdouzan, M.R.: A comparison of different haptic compression techniques. In: IEEE Intern. Conf. on Multimedia and Expo, ICME, vol. 1, pp. 657–660. IEEE (2002)
14. Eid, M., Cha, J., El Saddik, A.: Admux: An adaptive multiplexer for haptic–audio–visual data communication. IEEE Trans. on Instrumentation and Measurement **60**(1), 21–31 (2011)
15. Hamam, A., El Saddik, A.: Toward a mathematical model for quality of experience evaluation of haptic applications. IEEE Trans. on Instrumentation and Measurement **62**(12), 3315–3322 (2013)
16. Iwata, K., Ishibashi, Y., Fukushima, N., Sugawara, S.: Qoe assessment in haptic media, sound, and video transmission: Effect of playout buffering control. Computers in Entertainment (CIE) **8**(2), 12 (2010)
17. Suzuki, N., Katsura, S.: Evaluation of qos in haptic communication based on bilateral control. In: IEEE Intern. Conference on Mechatronics (ICM), pp. 886–891. IEEE (2013)
18. Isomura, E., Tasaka, S., Nunome, T.: A multidimensional qoe monitoring system for audiovisual and haptic interactive ip communications. In: IEEE Consumer Communications and Networking Conference (CCNC), pp. 196–202. IEEE (2013)
19. Allin, S., Matsuoka, Y., Klatzky, R.: Measuring just noticeable differences for haptic force feedback: implications for rehabilitation. In: 10th Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems, pp. 299–302. IEEE (2002)
20. Karadogan, E., Williams, R.L., Howell, J.N., Conatser Jr., R.R., et al.: A stiffness discrimination experiment including analysis of palpation forces and velocities. Simulation in Healthcare **5**(5), 279–288 (2010)
21. Silva, A.J., Ramirez, O.A.D., Vega, V.P., Oliver, J.P.O.: Phantom omni haptic device: Kinematic and manipulability. In: Electronics, Robotics and Automotive Mechanics Conference, pp. 193–198. IEEE (2009)
22. Cavusoglu, M.C., Tendick, F.: Multirate simulation for high fidelity haptic interaction with deformable objects in virtual environments. In: IEEE Int. Conf. on Robotics and Automation, ICRA, vol. 3, pp. 2458–2465. IEEE (2000)

An Efficient Content Searching Method Using Transmission Records with Wasted Queries Reduction Scheme in Unstructured Peer-to-Peer Networks

Yasuaki Ozawa and Shinji Sugawara

Abstract A lot of content searching methods applied to Peer-to-Peer (P2P) networks have been researched recently. However, especially in unstructured P2P networks, the amount of traffic grows when a flooding-based content searching is adopted because of a lot of wasted queries. Accordingly, we proposed a content searching method for unstructured P2P networks in our former research, in which each peer keeps query transmission records and guides queries properly so as to reduce wasted ones according to the records by canceling query transmissions to the peers with low content existing probability. In this paper, we improve the proposed method by adding wasted query reduction scheme which works while the usual flooding is executed, and reduce the network traffic without degradation of content acquisition rate. In addition, we discuss the availability of the newly proposed method with its evaluation results given by computer simulations.

Keywords Contents searching · Transmission record · Unstructured peer-to-peer

1 Introduction

A lot of content searching methods applied to Peer-to-Peer (P2P) networks have been researched recently. However, especially in unstructured P2P networks, the amount of traffic in the network grows enormously when flooding-based content searching strategy is adopted. Accordingly, we proposed a content searching method for unstructured P2P networks in our former research, in which each peer keeps query transmission records and guides queries properly so as to reduce wasted ones according to the records by canceling query transmissions to the peers with low content existing probability.

Y. Ozawa · S. Sugawara(✉)
Chiba Institute of Technology, 2-17-1 Tsudanuma, Narashino 275-0016, Japan
e-mail: shinji.sugawara@it-chiba.ac.jp

T.T. Zin et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 388,
DOI: 10.1007/978-3-319-23207-2_11

In this paper, we improve the prosed method by adding wasted query reduction scheme which works while the usual flooding is executed, and try to reduce the network traffic further without degradation of content acquisition rate. Then, we evaluate the efficiency of the newly proposed method by computer simulations, comparing to both former proposal and conventional flooding-based contents searching, from the viewpoint of network traffic and contents acquisition rate.

The remaining part of this paper is constructed as follows. Related works are introduced and proposed method is illustrated in Sections 2 and 3, respectively. Evaluation including some discussions is stated in Section 4 and at last, Section 5 concludes this paper.

2 Related Works

In this section, some typical conventional methods for content searching in peer-to-peer networks are introduced, which strongly related to the method proposed in this paper.

2.1 Breadcrumbs

Breadcrumbs [1] is a method to improve the efficiency of contents delivery for client-server based contents sharing systems by caching content items in network routers (nodes). In this method, load concentration to the content server can be reduced by deploying query guiding information on the nodes located on the content delivery pathways.

Figure 1 illustrates a typical behavior of Breadcrumbs. When a node located more closer to the server than another, the former node is called as an upstream node of the latter node. In an opposite manner, the latter node is called as an downstream node of the former node. In fig. 1, node D is upstream node of nodes A, B and C, and node A is downstream node of nodes B, C and D.

In this figure, node A already retrieved a certain content item from the server, and each node of B, C and D keeps query guiding information for the searching of the content item. The arrows possessed by the nodes mean the query guiding information showing their downstream nodes.

When node E requires a content item, which is the same item that node A requested in the past and still possesses now, node E sends a query for the item to node C and F. The query sent to node C is guided only to node B and then to node A, finally the item is found at node A and node E retrieves the item sent by node A reversely following the query's trail. Note that no query is sent from node C to D for this searching. By using query guiding, no wasted queries are sent along the trail after node C, and this method reduces more wasted queries for the discoveries of content items than flooding-based methods.

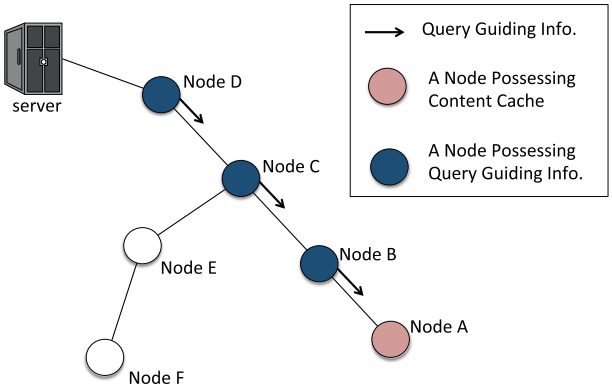


Fig. 1 Typical Behavior of Breadcrumbs.

2.2 APS

In Adaptive Probablistic Search (APS) [2], each peer keeps a table which consists of indices and index values. Indices are a list of adjacent peers which can be candidates of query forwarding destinations. Index value means adequacy of sending a query to the corresponding adjacent peer for finding the corresponding content item.

Basically, the higher index value an adjacent peer has, in the higher probability the adjacent peer is selected as a query forwarding destination. When a peer sends a query to a selected adjacent peer, the index values of all the adjacent peers in the table are uniformly reduced by a constant amount. After that, when the target content item is found by sending the query to the adjacent peer, the item is sent reversely following the query's trail, and an amount which is larger than the reduced amount is added only to the index value of the adjacent peer.

2.3 TTL Dynamic Control

TTL dynamic control [3] aims to improve the number of hit-queries (the query which reaches the peers possessing its target content item) per one query packet. In this method, each peer has a table which is similar to the one of APS and consists of adjacent peers' IP addresses and hit count which means the number of successful relaying to the peers possessing target content items.

Each query packet has TTL (Time To Live) value. The reducing amount of TTL for a relay from a peer to another is dynamically changed at each peer according to the hit count, in order to give larger TTL to the queries sent to the promising directions.

2.4 Conventional Method Using Content Transmission Records

This method is proposed by the group of one of the authors [4][5]. Originally, Breadcrumbs described above is for client-server system and accumulates query guiding information in the routers in the network. The concept is applied to the peer-to-peer network in this method, and query guiding information is used in both routers and peers simultaneously.

3 Proposed Method

In the conventional method introduced in section 2, the number of queries multiplied in the relaying peers which do not have TRs (Transmission Records) is not sufficiently reduced yet. Therefore in this paper, we improve the conventional method by applying a concept of TTL dynamic control method, introduced in section 2, in order to reduce the number of queries further.

3.1 Assuming Situation

Before illustrating the proposed method, we need to explain the assuming situation of contents sharing in this paper. Major assumptions are as follows.

- A number of peers are connected each other and make an unstructured peer-to-peer network for content sharing.
- Each peer has a certain limited capacity for possessing content items in its storage, and share the items with the other peers.
- Each peer can be disconnected from the network, and conversely, the disconnected peer can be connected again, any time with a certain probability.
- Each peer can be set TRs, that show the directions of sending content items from itself to adjacent peers just like Breadcrumbs introduced in section 2.
- When a peer requires a content item, basically the peer can find the item by flooding-based searching, and if the item is found, it is sent to the requesting peer reversely tracing the route of the query which reached the possessing peer.

3.2 Algorithm of Proposed Method

The algorithm of the method we propose is shown in the form of a combination of procedures each peer runs as follows. Because Breadcrumbs-like query control is executed basically based on the conventional method, precise explanations of some parts are omitted and we focus on the newly added part in this paper.

Content-Requesting Phase. In the proposed method, when a user requests a content item, the item is searched by flooding. In the case where a flooding based searching succeeds to find a target content item, each query-relaying peer memorizes the number of successes of target contents discoveries (i.e., N_{hit}) after relaying queries. The main contribution of this paper is to propose a method to reduce the number of wasted queries further by greatly decreasing queries' TTLs at the peers with small N_{hit}.

Parameters used in the explanation of the method are shown below. Note that the direction a TR points in to content-possessing peer on the route of content transfer is called "downstream," and the opposite direction is called "upstream." (Cf. Fig. 2)

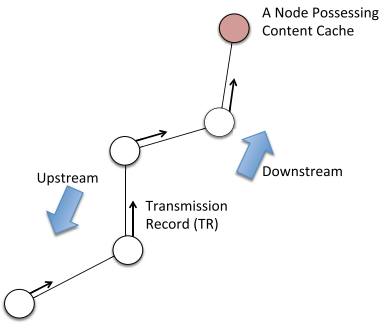


Fig. 2 Directions of "Upstream" and "Downstream."

- Elapsed time of possessing content n in the peer: $T(n)$
- Number of acquisition request from the other peers for content item n to the peer (initial value is zero): $N_R(n)$
- Availability of content n in the peer (initial value is zero): $F_R(n)$
- Threshold concerning the availability of a content item (an appropriate value is supposed to be given): R_{th}
- Number of TRs concerning content n in the peer (initial value is zero): $N_{TR}(n)$
- TTL (Time To Live) value each query has (an appropriate initial value is supposed to be given): V_{TTL}
- Number of hops a query advanced from requesting peer to content possessing peer: H
- Threshold concerning the querys' hops (an appropriate value is supposed to be given): H_{th}
- Threshold concerning the number of links to adjacent peers from a peer (an appropriate value is supposed to be given): L_{th}
- Number of successes of target content discovery resulting from relaying queries to adjacent peer x (initial value is zero): $N_{hit}(x)$
- Threshold concerning decrement of queries' TTLs (an appropriate initial value is supposed to be given): C_{th}

<Procedure begins>

1. When the content n the user requires is not in his/her peer, start the content searching. If $N_{TR}(n) = 0$ in the requesting peer, go to 2. Otherwise, go to 4.
2. Requesting peer or the peer received the query for the content n sends queries by flooding. If a query's V_{TTL} is zero, end this procedure in failure.
3. If the peer that receives the query has content item n, the procedure ends in success at the peer and switch to *Content-Discovery Phase* stated in the next part. Otherwise, if $N_{TR}(n) = 0$, once go to *sub-procedure*, explained later, to decrease V_{TTL}, and then, just after returning from the *sub-procedure*, go back to 2. In the case of $N_{TR}(n) \neq 0$, go to the next step. (Cf. Fig. 3)
4. At the peer, refer to the last updated TR concerning the requested content n and guide the query downstream according to the TR. During the guiding the query, V_{TTL} is not decremented. If the peer that receives the query has the requested content item n, the procedure ends in success at the peer and switch to *Content-Discovery Phase* stated in the next part. If the link to the peer to which the query is supposed to be guided is missing, go to 5. If the query reaches the most downstream peer, and the content n does not exist at the peer, go to 6.
5. Disable the TR of the downstream direction. Go to the next step.
6. Guide the query upstream according to the TRs until the query reaches the most upstream peer. During the transfer of the query, the peers on the path update their TRs, so as to reverse the guiding directions. If an upstream link is missing along the way, or the intended content item n does not exist in the most upstream peer, disable all the TRs concerning the content item n, and decrement $N_{TR}(n)$s in all the peers on the path, and then, go back to 2. If the intended content n is stored in the most upstream peer, the procedure ends in success at the peer and switch to *Content-Discovery Phase* stated in the next part.

<Procedure ends>

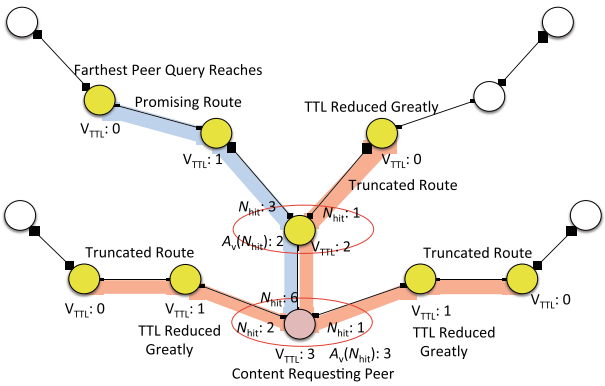


Fig. 3 TTL Control.

<Sub-procedure begins>

1. In the peer, if the number of links to the adjacent peer is smaller than L_{th}, send the query to all of the adjacent peers and go back to the main procedure shown above. Otherwise, go to the next step.
2. Refer to all of the adjacent peers for their N_{hit}s, and calculate the average $A_v(N_{hit})$.
3. Check each of the adjacent peers whether its N_{hit} is smaller than $A_v(N_{hit}) - C_{th}$. If it is true, go to the next step. Otherwise, decrement the query's TTL, send the query to the adjacent peer and go back to the beginning of this step 3. If all of the adjacent peers are checked, go back to the main procedure.
4. If $V_{TTL} > 2$, subtract 2 from V_{TTL} (i.e., $V_{TTL} \leftarrow V_{TTL} - 2$), send the query to the peer and go back to 3. Otherwise, give zero to V_{TTL} (i.e., $V_{TTL} \leftarrow 0$), send the query and go back to 3.

<Sub-procedure ends>

Content-Discovery Phase. After the discovery of a content item n, the following procedure is activated.

<Procedure begins>

1. In the peer at which the query found the intended content item n, increment $N_R(n)$, and from the peer, send the item to the requesting peer along the route the reached query advanced.
2. In each peer on the route, if there is the adjacent peer x which was sent the query which was eventually to discover the content, increment $N_{hit}(x)$.
3. In each peer on the route, if H is larger than or equal to the threshold H_{th}, create a TR concerning content item n if the peer does not have it yet, and set downstream direction toward the peer the target content item was discovered at. Otherwise, go to 6.
4. If $F_R(n)$ is larger than or equal to the threshold R_{th}, place the replica of content item n on content requesting peer, and increment the value of $N_R(n)$ at the peer.
5. All of the directions of TRs concerning content item n deployed along the route are reset, so as to guide the query to the newly placed replica from the next content request.
6. The value of $F_R(n)$ at the peer which possesses the content item n is recalculated by the following equation (1). W in (1) is set to an appropriate value as a weight.

$$F_R(n) = \frac{N_R(n) - W \cdot N_{TR}(n)}{\log T(n)} \tag{1}$$

<Procedure ends>

Table 1 Simulation Parameters.

Parameters	*Values*
Simulation Period (unit time)	5,000
Number of Simulation Runs	100
Number of Peers	5,000
Number of Contents	10
Number of Replicas per Each Content	10
TTL	2-7
Peers' Join-in and Drop-out Frequency (λ of Poisson distribution)	0.4-0.8
H_{th}	1-3
L_{th}	4
C_{th}	0.3

4 Evaluation

In this section, we show the effectiveness of the proposed method by computer simulations.

4.1 Measures of Evaluation

In order to compare the effectiveness of the methods, we use two measures, i.e., content acquisition rate and communication costs. The definitions of them are as follows.

- **Content Acquisition Rate:**
  Ratio of the number of successes of target content acquisitions in the total number of content requirements.
- **Communication Costs:**
  Summation of the numbers of hops the queries and messages are forwarded from a peer to another in the network when contents searching, making TRs, and renewing TRs.

4.2 Conditions for Computer Simulations

For the computer simulations, we set the parameters to the values shown in Table 1. Network topology is defined according to Barabási-Albert model [6].

Proposed method is compared with the other typical content searching methods such as Flooding and Conventional method explained in section 2 by the measures shown above.

4.3 Evaluation Results and Discussions

The results of the evaluations are as follows. Figure 4 shows the relationship between initial number of TTL and acquisition rate, and Fig. 5 illustrates the relationship between churn rate and communication costs. Note that the term "churn rate" here represents the frequency of peers' joining in or dropping out of the network, and actually means the arrival rate λ of Poisson distribution.

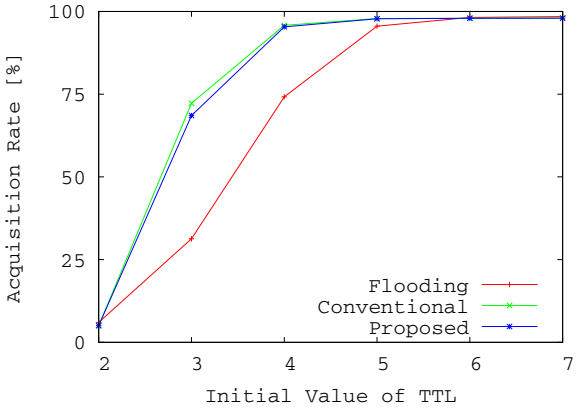


Fig. 4 Relationship between Initial Values of TTL and Acquisition Rates. (Churn Rate: 0.6)

In Fig. 4, the acquisition rates are getting larger and larger along the increase of initial number of TTL, and finally, they saturate and reach almost 100 % in any case of using one of the three methods. Proposed method achieves almost the same acquisition rate with conventional method and much higher rate than flooding in the condition where initial TTL is small. If sufficient values of initial TTL are given, acquisition rates reach 100 % by using any methods, however in that case, flooding sends a lot of wasted queries and this causes a heavy load in the network. Note that λ is set to 0.6 in this simulation (churn rate: 0.6). We evaluated in the situations where λ was changed to 0.4 and 0.8, and confirmed that the results were almost the same.

Fig. 5 illustrates the relationship between λ and communication costs. Proposed method successfully controls the cost better than conventional method and much better than flooding. Although the initial TTL was set to 7 in this simulation, we also evaluated in the case where the initial TTL was set to 3 and 5, and the result showed almost the same tendency.

Totally from the results of the simulations shown above, we can confirm that the proposed method achieves acquisition rate well at almost the same level with conventional method, and at the same time, controls communication costs better than the other methods.

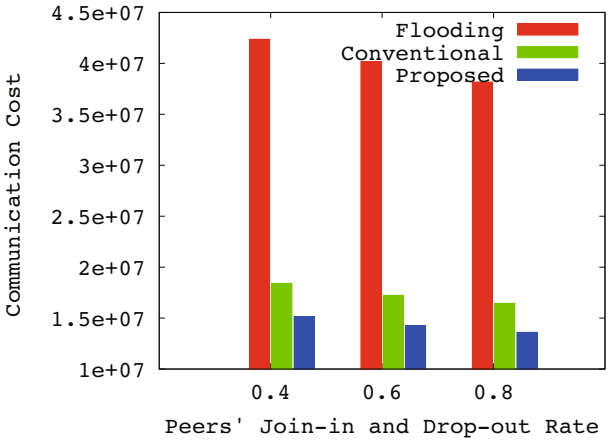


Fig. 5 Relationship between Churn Rates and Network Costs. (TTL: 7)

5 Conclusion

We proposed an improved contents searching method using transmission records in unstructured peer-to-peer networks. The main contributions in this paper is to control the number of wasted queries. In this method, when a query is not guided by Transmission Records (TRs) in a peer, the query is forwarded only to the adjacent peers with sufficient past successful experiences of finding target content items by forwarding queries to the peers.

According to the results of the evaluation, we confirmed that the proposed method achieved better performance than the conventional methods. As a future work, we will continue to improve the proposed method further mainly in the process of wasted queries omitting.

Acknowledgments This research was partially supported by JSPS KAKENHI Grant Number 25330129.

References

1. Rosensweig, E.J., Kurose, J.: Breadcrumb: efficient, best-effort content location in cache networks. In: Proc. IEEE INFOCOM 2009, pp. 2631–2635, April 2009
2. Tsoumakos, D., Roussopoulos, N.: Adaptive probabilistic search for peer-to-peer networks. In: Proc. ICS 2003, pp. 102–110, September 2003
3. Honma, T., Kawakami, H., Asatani, K.: A study on TTL dynamic control for efficient search in P2P networks. NS003-253, OCS2003-137, IEICE Technical Report, pp. 61–64, January 2004. (in Japanese)

4. Tomimatsu, T., Sugawara, S., Ishibashi, Y.: An efficient content searching method using transmission records in unstructured peer-to-peer networks. NS2012-86, IEICE Technical Report, pp. 37–42, October 2012. (in Japanese)
5. Tomimatsu, T., Sugawara, S., Ishibashi, Y.: Query guidance with transmission records for efficient content searching in unstructured peer-to-peer networks. In: Proc. IEEE 2013 International Communications Quality and Reliability (CQR) Workshop, May 2013
6. Barabási, A.L., Albert, R.: Emergence of scaling in random networks. Science **286**, 509–512 (1999)

Reliability Specification of Telecommunication Networks Based on the Failure Influence by Using Evolutional Algorithm

Pingguo Huang and Hitoshi Watanabe

Abstract For the first step of the reliability design of telecommunication networks, the reliability requirements should be determined. The method to determine the reliability requirements which takes account of the social loss caused by service outage has been established for telephone network in Japan. The feature of above method is to determine the reliability requirement of network element by solving the equation between the loss and reliability of the simple structure of networks, i.e. subscriber areas or transit lines. Therefore, the previous method cannot be applied for the general structure of networks, i.e. the NGN architecture consisted of various control equipment. This paper has investigated the possibility to determine the reliability requirement of network element evolutionally by allocating the traffic demand and the loss caused by failure to the concerned element.

Keywords Telecommunication networks · Reliability · Design · Requirement

1 Introduction

There are various countermeasures for improving network reliability. The Japanese telecommunication networks have experienced various failures and disasters and improved its reliability from the lessons learned by these experiences. Then, the backbone of Japanese telecommunication network has achieved the adequate reliability. To achieve high reliability under the reasonable cost, the reliability design is important. The process, i.e. to determine the reliability requirements, to evaluate the reliability of the possible alternatives, and to select the optimum solution is important. Especially to determine the reliability objective is the most important process.

P. Huang(✉) · H. Watanabe
Faculty of Engineering, Tokyo University of Science,
1-3 Kagurazaka, Shinjuku-ku, Tokyo 162-8601, Japan
e-mail: {huang,kwata}@ms.kagu.tus.ac.jp

T.T. Zin et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 388,
DOI: 10.1007/978-3-319-23207-2_12

About the reliability specification of telecommunication networks, a method taking account of the social loss caused by service outage has been established for telephone network in Japan [1]. The feature of this method is to determine the reliability requirement of network element by solving the equation related to the loss and reliability of the simple structure of networks, i.e. subscriber areas or transit lines. Therefore, this method cannot be applied for the general structure of networks, i.e. the NGN architecture consists of various control equipment. The authors have tried to establish the more general method for reliability specification [2], [3]. This paper describes the possibility to determine the reliability requirement of network element evolutionally by allocating the traffic demand and the loss caused by failure to the concerned element.

2 The Present Method

Because telecommunication networks have complex and spreading structures, there are various reliability measures and to select the appropriate measures for specification is difficult. Considering these situations, the telephone network is modeled as the network which consists of the subscriber areas and transit network (Fig. 1) and the following three measures are used for reliability specification. The first is the end to end reliability. This is the probability that the path between the concerned pair of users in network is available. The second is the unavailability requirements of subscriber exchanges considering system size. The third is the area to area reliability as the function of degradation rate of transmission capacity.

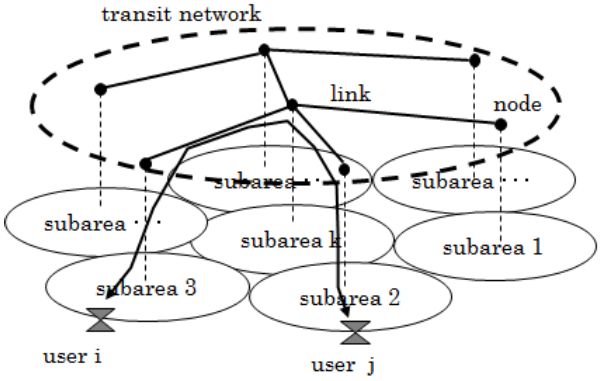


Fig. 1 The model of telephone network.

Here, the reliability requirements of subscriber exchanges are determined as following process by using the concept of social loss caused by the service outages [4], [5]. A failure of subscriber exchange yields the complete service outage of the area. It is thought that the area suffered a certain loss per unit time. Therefore, it can be defined the total loss per unit time caused by the failure. The loss is explained

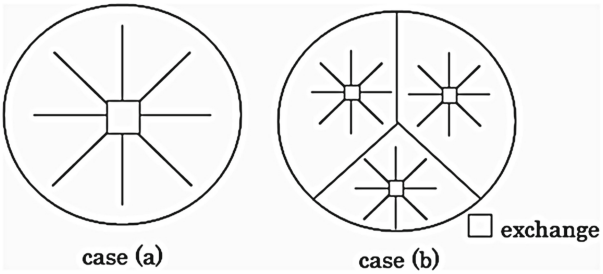


Fig. 2 Service by exchanges with different sizes.

as L(x). Let consider the two cases for serving telecommunication as shown in Fig. 2. One exchange with size x serves total area in the case (a) and n exchanges with size x/n serve for same area in case (b).

From the view point of users, the total expected loss of whole area per unit time should be equal in these cases, because the way to service is the matters of provider and the users have nothing to do it. Then, the equation about the relation of expected loss of two cases is obtained as formula (1)

$$U(nx)L(nx) = nU(x/n)L(x/n) \tag{1}$$

Here, $U(\ )$: Unavailability
$L(\ )$: Loss

This is a functional equation of U(x). Therefore, U(x) is solved as follows.

$$U(x) = kx/L(x) \tag{2}$$

Here, k is the constant

This formula enables to determine reliability requirement as a function of its size by using the concept of social loss caused by service outages. Here, the meaning of social loss should be concerned. There are some understandings. The first is a kind of thought experiment. The literature [4] thinks that the users may go to the outside of subscriber area to communicate if the service stops. Then, the total moving distance of users in that area represents the social loss. Therefore, the following formula is obtained, because the average distance is proportioned to radius of the area and number of users is proportioned to square of radius.

$$L(x) = cx^{1.5} \tag{3}$$

Here, c is constant

In this case, the unavailability objective of system with size x is proportioned to -0.5 power of x. There are other investigations about the social loss. Literature [6] analyzed the field data of outage in America and yielded a social loss as a function of

outage duration and number of users affected by outage. Literature [7] analyzed the Japanese field data in details not only for telephone service but also for other telecommunication services. These results reinforce that the social loss is proportioned to 1.5 power of number of affected users.

In transit network, the reliability requirements are expressed as the function of transmission capacity between two areas. The influence of failure of facilities appears to the degradation of transmission capacity in transit networks. For example, the two area connected by two routes (Fig. 3) has the three levels of degradation. Those are complete cut off, half cut off and normal state. Therefore, to determine the reliability objective as the function of degradation rate of transmission circuits is reasonable. In actual design, the function is specified as like as a step-wise graph and the network hierarchy is reflected to the severity of requirements [1].

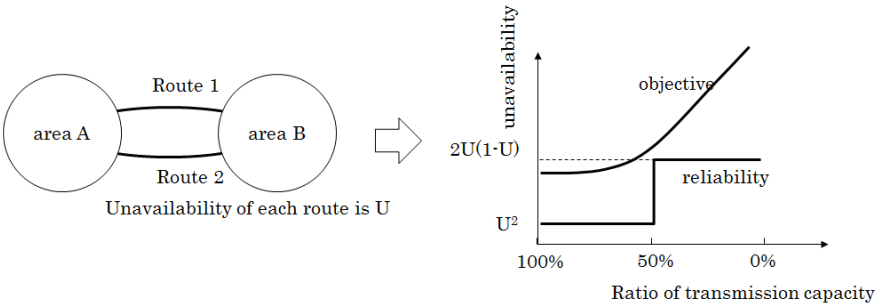


Fig. 3 Reliability requirements of transit networks.

3 Reliability Design of Signaling Network

The call processing in telephone network is carried out by using the call control signals. Currently these signals are transmitted by the network which is constructed separately with the voice transfer network. This is called Common Chanel Signaling Network(CCS Network). The CCS network is very important, because its failure must yield the large service outage. Therefore, the question how to keep the high reliability of CCS network with reasonable cost occurred. The current redundancy is the complete double and those two networks are constructed separately in the level of physical routes and only one network has the adequate transmission capacity. Although this structure may be thought too much reliability, this is the reasonable in practical sense. Because the transmitted information amount by CCS network is much smaller than voice transfer network, the complete redundancy does not increase so much cost in total.

However, the appropriateness of this simple design method is questionable in the future networks, such as NGN, in which the control network and transmission network cannot be separated clearly and the amount of transmitted information by

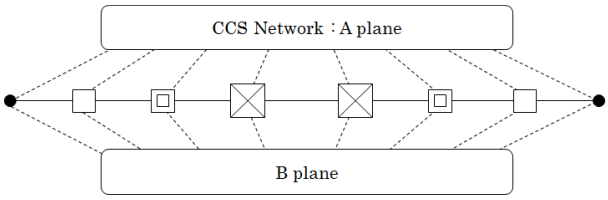


Fig. 4 The CCS Network.

control network become to be large. In these conditions, the new design method is required. Then, the problem is modelled as follows.

4 Model and Problem

- The network offers a telecommunication service to an area with distributed users with homogeneous density. This area is divided into N subareas.
- Each subarea has its representative node, which is supposed as a subscriber exchange.
- The number of users in subarea k is n_k. These users are connected to their representative node k (k= 1, ... , N).
- The number of user in whole areas is NZ.
- There are (M-N) nodes except for above N nodes.
- The transit network consists by connecting those M nodes by links with arbitrary topology.
- In normally, there are two paths between two nodes. A path consists of the series of nodes and links. The communication between two nodes is available only when at least one of these paths is available. Although there are various methods to set two paths between two nodes, the following method is assumed here.
  (1) The first path is the shortest route from starting node to destination node.
  (2) The second path is the shortest route which is consist by the remained links after eliminate links used by the first route. If there is not such route, only the first path is considered in analysis.
- S servers are distributed on the above M nodes. Each node has at most one server.
- The necessary and sufficient condition for communicating two nodes is that there is at least one path between two nodes and at least one server can be reached from starting node.
- Nodes and links have their own failure probability. The failure probabilities are expressed as unavailability. The failures are occur independently.

The problem is to determine the unavailability of nodes and links in accordance with some kind of criteria. The followings are candidate criterion. The first is to keep the reliability degradation caused by control network enough smaller than it caused by the failures of voice transfer network. The second is to equalize expected loss of users in overall network. This paper investigated the second way.

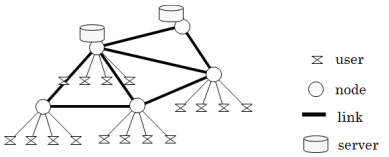


Fig. 5 The model of networks.

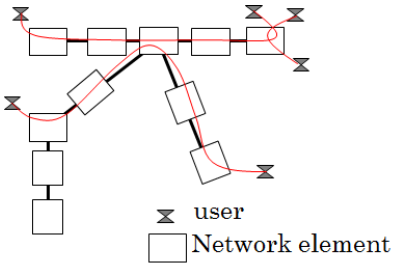


Fig. 6 The allocation of demand and loss.

From formula (2), the following formula is obtained.

$$U(x)L(x)/x = k \tag{4}$$

This means that the formula (2) can be recognized as the method to equalize the loss of user in network. In other word, that realizes the equalization of user's loss and builds a network which is well balanced reliability. Therefore, also in this paper, the second criteria will be investigated.

5 Determination of Unavailability of Elements

5.1 The Basic Procedure

The left-hand side of formula (2) is obtained by dividing the number of users in the area by the whole expected loss. Then, we apply this formula to determine the unavailability of network element also in general structured networks as shown in Fig. 5. Of course, this way is not strict. Because, in the general network, an element takes the part of different combination of communication, the relationship between unavailability and loss is not described as simple as the formula. However, we apply this method for obtaining the first order approximation. The procedures are summarized as follows.

Procedure 1 (When only 1 path is available)
Provide the two parameters LX_k and Lf_k. Let $\wedge_i$ be the total originating rate of communication from user i to another users. Pick out the elements which are necessary to communication all combination of i and j (Fig. 6). Let m be the number of necessary elements. The above parameters LX_k and Lf_k are incremented as following way.

$$LX_k = LX_k + \lambda_{ij} L_{ij} \tag{5}$$

$$Lf_k = Lf_k + \lambda_{ij}/\Lambda_{ij} \tag{6}$$

Here, λ_{ij} is the originating rate from user i to j and L_{ij} is the loss when the communication stops. In another word, this process is called as an allocating process

of telecommunication demand and loss to passed elements. After the allocation, the unavailability is obtained by the next formula.

$$U_k = cLf_k / LX_k \tag{7}$$

Here, c is the constant

Procedure 2 (When two paths are available)
The elements on the path between node i to j are divided into two classes. The first is the elements which cause the stop of communication when only the element fails. Let call such element as “one failure element". The second is that which cause stop of communication when the failure occur in said element and other element at same time. Let call such element as “two failures element". An element which is once recognized as “one failure element" is also recognized in other occasion as “two failures element" according with the difference of communication pairs. Then, 4 parameters are provided and the 2 parameters are for “one failure element" and another 2 are for “two failures element". The allocation of “one failure element" is carried out by formula (5) and (6). The allocation of “two failures element" is carried out by the following formulae (8) and (9).

$$LY_k = LY_k + \lambda_{ij} L_{ij} \tag{8}$$

$$Lg_k = Lg_k + \lambda_{ij} / \Lambda_i K \tag{9}$$

Here, K is the number of elements relates to the concerned communication.

After the above allocation, the unavailability is obtained as follows. Let consider the effects of the failure of an elements with unavailability U. The effect is expressed as the following formula.

$$L = U \Sigma L_i + U^2 \Sigma L_j \tag{10}$$

Here, L_i is the loss suffering to user i yielded by the roll of “one failure element" and L_j is the loss yielded by the roll of “two failures element". Because the ordinary unavailability is much smaller than 1, U^2 is much smaller than U. Therefore, the second term of right-hand side is rather smaller than first term. Therefore, when the considered element plays both kind of element, the second term can be neglected. Therefore, in such case, the unavailability is obtained by the formula (7).

On the contrary, when the concerned element plays only as “two failures element" for any communication, the unavailability is obtained by the following formula.

$$U^2 \Sigma L_j / x = const \tag{11}$$

Then, U is obtained as follows.

$$U = C(LY_k / Kg_k)^{1/2} \tag{12}$$

Here, C is the constant.

5.2 Evaluation Method of Expectation of Loss

As this paper supposes that failures occur independently, the expectations of loss of each user are evaluated as follows. At first, the state whether elements are up or down are listed up. In the system with N elements, the number of states is 2^N. This becomes to be rather large number in actual system. However, if the states are restricted in which the largest number of failed element is some number, the number of concerned state is not large. In actually, as the probability of failure is rather small, this approximation is reasonable. Secondly, the probability of each state is calculated. Thirdly, the propriety of communication between whole combinations of users is examined and loss of each user is cumulated.

5.3 Increase the Precision by Evolutional Process

The process mentioned in section 5.1 is only approximation method. Then, this paper proposes the evolutional process to increase the accuracy by repeating the allocation process. The concrete process is described as follows.

Let note the following characteristics of the expected loss of each user. Let $\{ U_1, U_2, \ldots, U_M \}$ be the unavailability of element and $\{ L_1, L_2, \ldots, L_{NZ} \}$ be the expected loss of each user. L_i (i = 1, ... , NZ) are increasing functions of U_k (k = 1, ... , M). Therefore, if the unavailability of the element related to the communications of user which has large expected loss is reduced, the expected loss of the user becomes to be small. On the contrary, the allocating method mentioned in section 5.1 is the method to reduce the unavailability of element with importance to be small. Therefore, the allocating method has the possibility to revise unavailability of elements. Concretely, the additional allocation is carried out by the following process. At first, specify the subarea to be reduced or increased its expected loss. In case to reduce the user loss, only the denominator is incremented in allocation process. In case to increase, only the numerator is incremented.

However, this revision may affects to other users and there is the possibility of degradation of total balance in network. Therefore, we have investigated the effectiveness of this method by some network models.

6 Evaluation Examples

6.1 Necessity of Reliability Specification

Fig. 7 is the analyzed area. The area is divided into 6 subareas. The users are distributed in this area with homogeneous density. The loss means the distance between two users with communication.

Fig. 8 shows the expected loss of whole users when the unavailability of node are equal and unavailability of link is 0. It shows that the expected loss of user in subarea 6 is much greater than other areas. The area become to be greater, the combination and distance is become to be greater in nonlinearly. To balance these

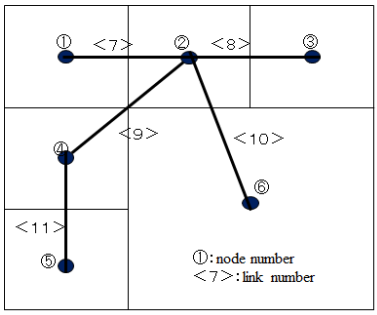


Fig. 7 The area concerned.

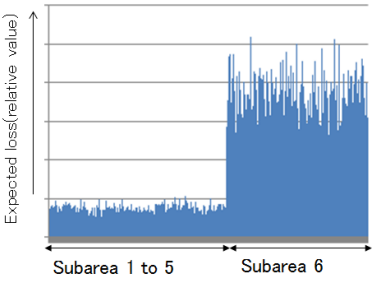


Fig. 8 Expected loss (1).

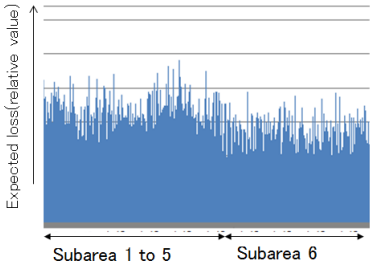


Fig. 9 Expected loss (2).

two tendencies, set the reliability objective of greater exchange be smaller is the reasonable solution. Then, Fig. 9 shows the expected loss of users under the condition that the unavailability of exchange is proportioned to -1.5 power of the number of user in subareas. The expected losses of users are well equalized. Fig. 10 shows the expected loss of users when the unavailability of elements are obtained by the method of allocating. Here, table 1 shows the unavailability obtained by our method. This shows that our proposed method has the possibility to determine the reliability objective automatically according to the importance of elements.

6.2 The General Network with Control Servers

There is no control server and only one path between nodes in the example of section 6.1. The next example is the network with control server and has rather general network topology and which has two paths between nodes (Fig. 11).

The unavailability of nodes and links are determined by the method in section 5.1 and the expected losses of users are evaluated. In these graphs, the average expected loss in each subarea is shown. Fig. 12 (a) shows the result of first allocation. In this result, the loss of subarea 1 is large. Therefore, in the second allocation process,

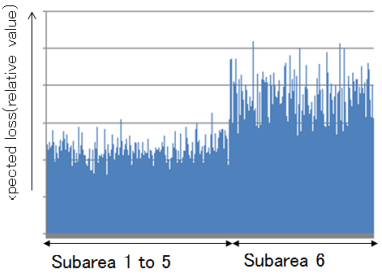


Fig. 10 Expected loss (3).

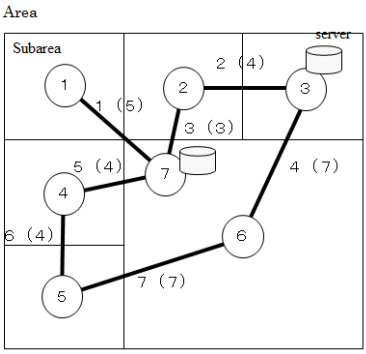


Fig. 11 The network with servers.

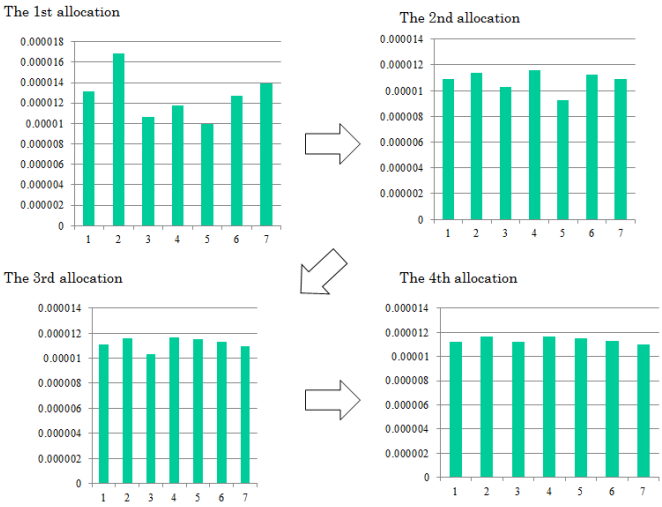


Fig. 12 Equalization of user loss by evolutional algorithm.

the allocation is carried out to reduce unavailability of elements related to the communication of subarea 1. Eventually, the allocation is repeated four times and the expected loss have become to be rather equalized. This shows that possibility of our proposed method.

Next case is the network with changing the location of servers. As shown in table 2, reliability objective is severer than Fig. 12. This is concluded from that Fig. 13 has only one server.

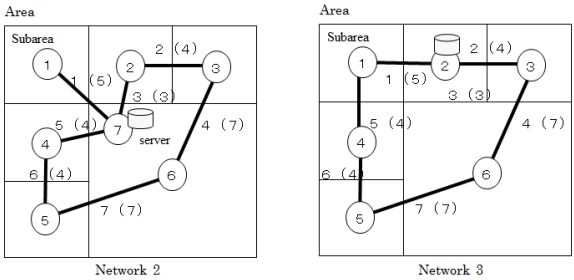


Fig. 13 Networks concerned.

Table 1 Unavailability of Element.

NE	Unavailability
1	3.58×10^{-5}
2	3.57×10^{-5}
3	3.78×10^{-5}
4	3.23×10^{-5}
5	3.93×10^{-5}
6	9.02×10^{-6}
7	0.048
8	0.05
9	0.03
10	0.019
11	0.062

Table 2 Unavailability of Element.

NE	Network 1	Network 2	Network 3
1	6.21×10^{-5}	6.41×10^{-5}	0.00017
2	0.000201	0.000161	0.000199
3	0.00024	0.000124	0.0000167
4	0.000312	8.96×10^{-5}	0.000126
5	0.000165	9.18×10^{-5}	0.000124
6	8.03×10^{-5}	3.48×10^{-5}	7.78×10^{-5}
7	6.72×10^{-5}	4.14×10^{-5}	9.81×10^{-5}
8	6.72×10^{-5}	7.24×10^{-5}	0.000101
9	6.81×10^{-5}	0.000101	8.02×10^{-5}
10	9.37×10^{-5}	0.000123	8.32×10^{-5}
11	5.98×10^{-5}	8.47×10^{-5}	7.06×10^{-5}
12	0.000123	5.54×10^{-5}	5.49×10^{-5}
13	9.6×10^{-5}	9.35×10^{-5}	0
14	6.66×10^{-5}	3.37×10^{-5}	3.37×10^{-5}

7 Conclusion

This paper proposed a new method to determine the reliability objective of telecommunication networks. The feature is to allocate communication demand and loss to network elements. This method has the possibility to apply for general network structure, such as the network which is not separate clearly transmission layer and control layer. This paper investigated the effectiveness only for restricted models. We would like to investigate in more general network models.

This work was supported by JSPS KAKENHI Grant Number 25350460.

References

1. Iwasaki, S., Tazaki, K.: For your good understanding about NTT Network, The Telecommunications Association (1987). (in Japanese)
2. Watanabe, H., Huang, P.: A consideration of the network reliability model for disaster risk reduction, IEICE Technical Report, R2014–68, pp. 19–24 (2014). (in Japanese)
3. Watanabe, H., Huang, P. : A consideration of the method to determine reliability requirement of networks based on the loss caused by service outage, IEICE Technical Report (2015). (in Japanese)
4. Nojo, S.: Evaluation of reliability equivalence considering failure-scale. IEICE Trans. on Fundamentals **J–64(A)**(1), 9–14 (1981)
5. Nojo, S., Watanabe, H.: Reliability Specification for Communication Networks Based on the Failure-Influence, IEEE Globecomf 1987, Tokyo, pp. 1135–139 (1987)
6. Tollar, E.S., Bennett, J.M.: Network outage impact measures for telecommunication, Symposium on Computers and Communications, Alexandria (1995)
7. Funakoshi, H., Matsukawa, T., Watanabe, H.: An Analysis of Social Influence of Outages for Telecommunication Network. IEICE Trans. on Commun. **90–B**(4), 370–381 (2007)

Trade-off Relationship Between Operability and Fairness in Networked Balloon Bursting Game Using Haptic Interface Devices

Mya Sithu, Yutaka Ishibashi, Pingguo Huang and Norishige Fukushima

Abstract This paper investigates the trade-off relationship between the operability of haptic interface device and the fairness between players by carrying out subjective and objective QoE (Quality of Experience) assessments in a networked balloon bursting game. In the game, two players burst balloons in a 3D virtual space by using haptic interface devices, and they compete for the number of burst balloons. As a result, we confirm that there exists a trade-off relationship between the operability and fairness; that is, if we try to improve the fairness, the operability is degraded when the network delays are different from terminal to terminal; if we try to improve the operability, the fairness is damaged. We also find that the contribution of the fairness is larger than that of the operability to the comprehensive quality (i.e., the weighted sum of the operability and fairness). Assessment results further show that the output timing of terminals should be adjusted to the terminal which has the slowest output timing to maintain the fairness when the difference in network delay between the terminals is large. In this way, the comprehensive quality at each terminal can also be maintained as high as possible.

Keywords Networked real-time game · Virtual environment · Balloon bursting game · Haptic interface devices · Network delay · Quality of experience · Operability · Fairness

1 Introduction

Players can achieve a high sense of immersion by using haptic interface devices in networked real-time games [1]-[4]. When the players play such games over a

M. Sithu(✉) · Y. Ishibashi · N. Fukushima
Graduate School of Engineering, Nagoya Institute of Technology, Nagoya 466-8555, Japan
e-mail: s.mya.492@stn.nitech.ac.jp, {ishibasi,fukushima}@nitech.ac.jp

P. Huang
Faculty of Engineering Division II, Tokyo University of Science, Tokyo 162-8601, Japan
e-mail: huang@ms.kagu.tus.ac.jp

T.T. Zin et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 388,
DOI: 10.1007/978-3-319-23207-2_13

network like the Internet, the consistency (e.g., the positions of an object at terminals are the same) of objects in a 3D virtual space at the players' terminals may be disturbed owing to the network delay, delay jitter, and packet loss.

To keep the consistency high at the terminals, we can employ the local lag control [5], which buffers the local information for a constant time called the local lag according to the network delay from the local terminal to the other terminal. Thus, the interactivity is degraded. The operability of haptic interface device is also deteriorated [6]. Moreover, when the difference in network delay between the terminals is large; the fairness among the players is largely damaged [7]. Note that the fairness in the game means that the same condition is provided to all the players [8]. For example, in a networked real-time game when two players play, when the network delay from a terminal to the other terminal is large, and that in the opposite direction is small, the operability is seriously degraded only at the terminal with the larger network delay under the local lag control. This leads to unfairness between the players.

To maintain the fairness high, we can use the adaptive Δ-causality control [9], which also employs the local lag control. The adaptive Δ-causality control generally sets the local lag at each terminal to the maximum network delay among the terminals. Therefore, when the maximum network delay is large, the operability of the haptic interface device may seriously be degraded, but the fairness is maintained high. Based on this relationship of the operability and fairness, we can say that there is a trade-off relationship between the operability and fairness as shown in [3] and [9], where the influence of network delay on the operability and fairness for hard objects are investigated. It is also important to carry out the investigation for soft objects. This is because the influences for soft objects may be different from those for hard objects.

In [6], a balloon bursting game in which two players burst balloons (i.e., soft objects) using haptic interface devices is dealt with in a 3D virtual space. The influence of the network delay on the operability of haptic interface device is investigated by QoE (Quality of Experience) [10] assessment. Assessment results show that the operability depends on the local lag, and the characteristics of soft objects change when the local lag is large; that is, the balloon becomes harder and more slippery. The change in characteristics of soft objects is different from that of hard objects; the hard ones become heavier as the network delay increases in [11]. The allowable range of local lag is around 150 ms; that is, the deterioration in the operability is allowable when the local lag is smaller than or equal to about 150 ms. In [7], QoE assessment on the fairness between players is carried out for the balloon bursting game [6]. In the game, the players compete with each other for the number of burst balloons. Each player who bursts more balloons than the other player wins the game. Assessment results show that the fairness between players is mainly dependent on the difference in network delay between the terminals, and the allowable range of the absolute difference in network delay between the players is within around 75 ms. In [6] and [7], QoE assessments on the operability and fairness are carried out separately. It is not clear how the operability

and fairness are related to each other. Therefore, it is important to investigate the relationship in detail by QoE assessment in which we should also assess the comprehensive quality (i.e., the weighted sum of the operability and fairness) to examine which QoE (operability or fairness) has larger contribution to the comprehensive quality. Also, we should clarify how much local lag should be set at each terminal to maintain the comprehensive quality as high as possible. For example, if the fairness has larger contribution to the comprehensive quality, the comprehensive quality may be maintained high by adjusting the output timing of the terminals to the terminal which has the slowest timing as in the adaptive Δ-causality control.

Therefore, in this paper, we investigate the trade-off relationship between the operability and fairness by carrying out subjective QoE assessment on operability, fairness and comprehensive quality in the balloon bursting game [6], [7]. The reason why we employ the game is that we here examine the relationship between the operability of haptic interface device [6] and the fairness among the players [7] for soft objects. We also perform objective QoE assessment at the same time as the subjective assessment. We further investigate the relationship between subjective and objective results.

The remainder of this paper is organized as follows. Section 2 describes the balloon bursting game. Assessment environment is explained in Section 3. Assessment results are presented in Section 4, and Section 5 concludes the paper.

2 Balloon Bursting Game

The system configuration of the balloon bursting game is shown in Fig. 1, where there are two terminals (*terminals 1* and *2*), each of which has a PC with a display, a haptic interface device (Geometric Touch [12]), and a headset. Each of two players (*players 1* and *2*) bursts balloons with his/her haptic interface device in a 3D virtual space. The two players compete with each other for the number of burst balloons. As shown in Fig. 1, we employ four balloons for simplicity in the virtual space. Each player employs his/her haptic interface device to move the virtual stylus in the 3D virtual space. In Fig. 1, the virtual styli are placed at the initial position. When the player touches the balloon with the tip of the stylus, the reaction force is perceived through the haptic interface device; he/she can feel the softness of the balloon. The balloon is distorted when the player pushes it with the stylus. If he/she pushes it strongly, the balloon is largely distorted, and it is burst and disappeared. Then, he/she hears a sound of bursting it via the headset.

In this paper, player 1 bursts two blue balloons alternately on the left side of the virtual space, and player 2 bursts two pink balloons on the right side as in [7]. This purpose is to avoid the situation of trying to burst the same balloon simultaneously at the two terminals for simplicity. Before the start of the game, the players stand ready by placing their styli at their respective initial positions (see Fig. 2 (a)). The players start to burst the balloons when “START” message is displayed on the screen (see Fig. 2 (a)) and a buzzer sound is output. During the game, the numbers

of balloons burst by the two players are displayed on the screen (see Fig. 2 (b)). The players stop the game when "GAME OVER" message appears on the screen 30 seconds after the beginning of the game (see Fig. 2 (c)). The buzzer sound also alerts the players to stop the game at that time. A player who bursts more balloons than the other player wins the game. When a balloon is burst and disappeared, a new balloon automatically appears at the location of the burst balloon. Both players try to burst their respective balloons from the front side of the balloon as fast as they can.

The reaction force applied to the haptic interface device is generated by the haptic rendering engine [13], which uses the object shape and material properties such as s tiffness and friction for calculation of the reaction force. The force applied to a balloon when the player pushes the balloon with the stylus is equal to the reaction force against the player. The player feels larger reaction force as the penetration depth of the stylus becomes larger; the volume of the balloon decreases in this paper. The penetration depth of the stylus is the distance from the surface of the balloon to the tip of the stylus. There may be several methods of judgment of bursting a balloon. In this paper, we use a method in which a balloon is burst when the volume of a balloon reaches a threshold value as in [6]. We set the threshold value to 90% of the initial volume of the balloon in our assessment. The size of the balloon in the virtual space can be changed. In this paper, the radii of three dimensional axes (x, y, and z) of the balloon are 1.1, 1.5 and 1.1, respectively (see the virtual space in Fig. 1), where we assume that the length of the stylus is 1.0.

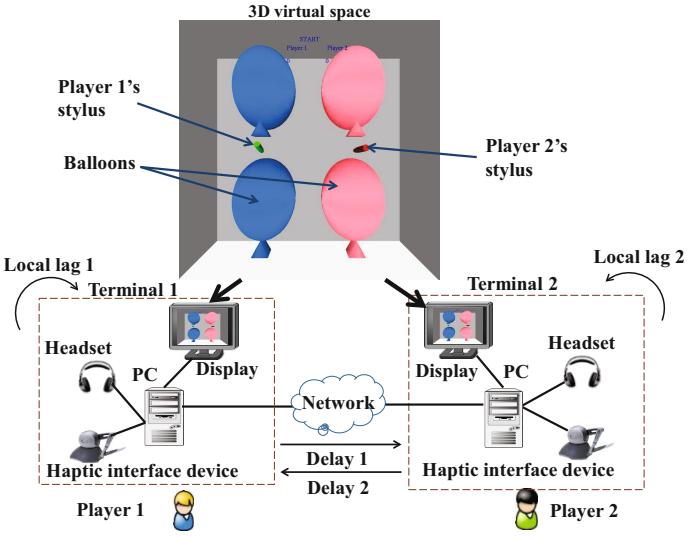
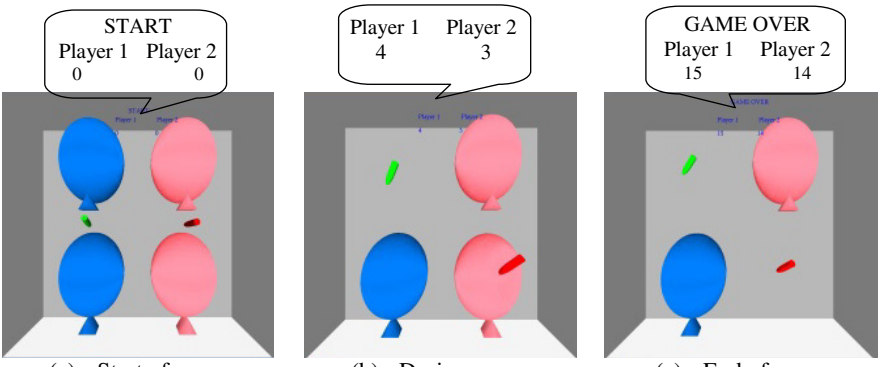


Fig. 1 System configuration of balloon bursting game.



(a) Start of game (b) During game (c) End of game

Fig. 2 Displayed images in virtual space.

3 Assessment Environment

3.1 Assessment System

In our assessment system, the two terminals are connected to each other via a network emulator (NIST Net [14]) which is used instead of the network shown in Fig. 1. The network emulator generates an additional constant delay for each packet transmitted between the terminals. Note that the network delay jitter can be absorbed by buffering under media synchronization control such as the Virtual-Time Rendering (VTR) algorithm [15]; we here take account of the jitter by including the buffering time in the constant delay as in [16]. We call the constant delay from terminal 1 to terminal 2 *delay 1*, and that from terminal 2 to terminal 1 *delay 2* (see Fig. 1). We call the local lag at terminal 1 *local lag 1* and that at terminal 2 *local lag 2*.

3.2 Assessment Methods

Before the assessment, each pair of subjects played the balloon bursting game for three times to get used to the game on the condition that delays 1 and 2 are set to 0 ms; that is, the same condition is provided to the pair. By practicing, each subject knows how to burst a balloon by using a haptic interface device. We carried out the assessment by setting delay 1 to 0 ms or 200 ms, and setting delay 2 to 50 ms, 100 ms, 200 ms, 300 ms, or 500 ms. We carried out QoE assessment with 16 subjects (males and females) whose ages were between 20 and 30.

Assessments in which delay 1 is set to 0 ms and 200 ms are referred to as *cases 1* and *2*, respectively. Local lag 1 is changed from 0 ms to 500 ms at intervals of 50 ms, and local lag 2 is set to the same value as delay 2. In each case, the order of combinations of delays and the local lags were changed in random order for the

pair. It took 30 seconds for each stimulus. After each stimulus, each subject was asked to base his/her judgment about the operability, fairness, and comprehensive quality in terms of wording used to define the five-grade impairment scale (5: Imperceptible, 4: Perceptible, but not annoying, 3: Slightly annoying, 2: Annoying, 1: Very annoying) [17]. In each stimulus, if the pair obtained almost the same results about victory or defeat as those in the practice, the pair regarded the fairness as high and valued the score at 5. The comprehensive quality is the weighted sum of the operability and interactivity; thus, the comprehensive quality is the most important. Each subject gave a score from 1 through 5 to each stimulus. By averaging the scores of all the subjects, we obtained *Mean Opinion Score* (MOS) [17]. We also adopted the number of burst balloons as an objective assessment measure. The total assessment time for each case was about two hours per a pair.

4 Assessment Results

4.1 Subjective Assessment Results

4.1.1 Operability and Fairness

We show the MOS values of operability at terminals 1 and 2 as a function of local lag 1 in Figs. 3 and 4, respectively. In Figs. 5 and 6, we also plot the MOS values of fairness at terminals 1 and 2, respectively. In the figures, we show only assessment results of case 1. We do not show results of case 2 since they had similar tendencies to those in case 1. The 95% confidence intervals are also plotted in the figures.

In Fig. 3, we see that the MOS values of operability for all the values of delay 2 at terminal 1 decrease as local lag 1 increases. This is because the local information at terminal 1 is buffered for a time of local lag 1; thus, since the interactivity is degraded, each subject feels that the balloon becomes harder and more slippery, and it is difficult to burst the balloon. From Fig. 4, we find that the MOS values at terminal 2 hardly depend on local lag 1, and they depend on mainly delay 2 or local lag 2. This is because the local information at terminal 2 is buffered for a time of delay 2.

In Figs. 5 and 6, we notice that when delay 2 is 50 ms, the MOS value of fairness hardly depends on local lag 1. For the other values of delay 2, the MOS values increase as local lag 1 increases. This is because each subject feels fairness strongly when the absolute difference of local lags 1 and 2 becomes smaller.

By comparing Figs. 3 and 5, we confirm that there is a trade-off relationship between the MOS value of operability and that of fairness at terminal 1; that is, the MOS value of operability becomes smaller when that of fairness becomes larger. Also, from Figs. 4 and 6, we note that the MOS value of operability at terminal 2 does not change when the MOS value of fairness becomes larger.

To clarify the optimum local lag values at both terminals, we have also calculated the average MOS values of two terminals, which are shown in Figs. 7 and 8. In Fig. 7, we see that the average MOS value of operability for two terminals decreases as local lag 1 becomes larger for each value of delay 2. In the figure, we

also find that delay 2 of 50 ms has the highest MOS value, and delay 2 of 500 ms has the lowest one for each value of local lag 1. In Fig. 8, we notice that when delay 2 is 50 ms, the average MOS value of fairness for two terminals hardly depend on local lag 1. For the other values of delay 2, the average MOS values increase as local lag 1 becomes larger. From the average MOS values of two terminals, we further confirm that the trade-off relationship exists between the MOS value of operability and that of fairness.

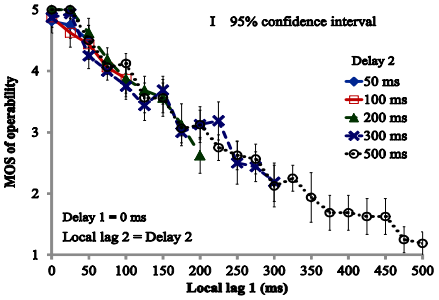


Fig. 3 MOS of operability at terminal 1.

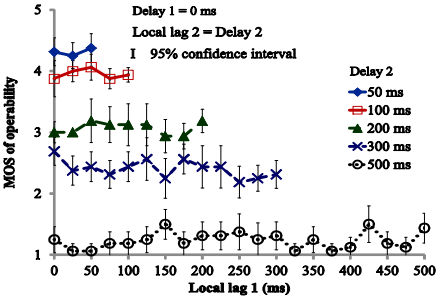


Fig. 4 MOS of operability at terminal 2.

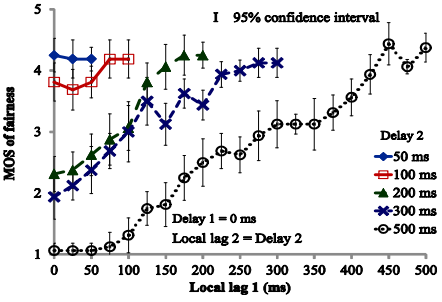


Fig. 5 MOS of fairness at terminal 1.

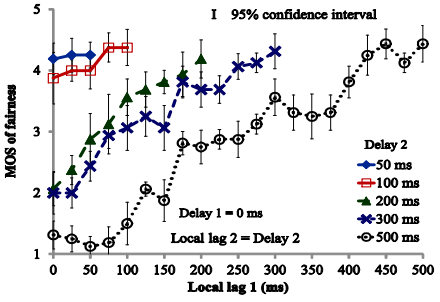


Fig. 6 MOS of fairness at terminal 2.

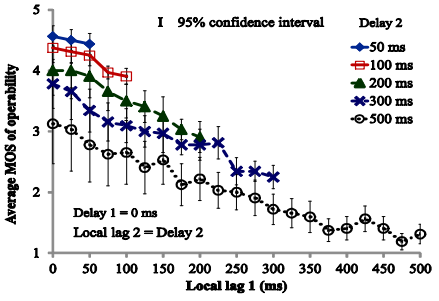


Fig. 7 Average MOS of operability for two terminals.

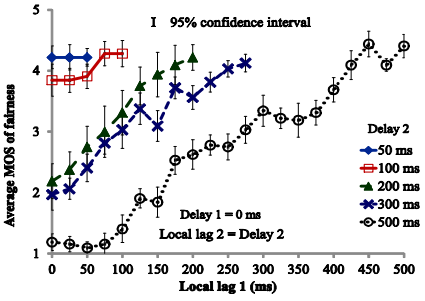


Fig. 8 Average MOS of fairness for two terminals.

4.1.2 Comprehensive Quality

We show the MOS values of comprehensive quality at terminals 1 and 2 in Figs. 9 and 10, respectively. The average MOS value of comprehensive quality is also plotted in Fig. 11.

Fig. 9 reveals that the MOS values of comprehensive quality for all the values of delay 2 at terminal 1 hardly depend on local lag 1. In Fig. 10, we find that the MOS value hardly depends on local lag 1 when delay 2 is 50 ms or 100 ms. For the other values of delay 2, the MOS values increase as local lag 1 becomes larger. From Fig. 11, we note that the average MOS value for delay 2 of 50 ms hardly depends on local lag 1. For the other values of delay 2, the average MOS values slightly increase as local lag 1 becomes larger. When delay 2 is 100 ms, the optimum value of local lag 1 is 100 ms. When delay 2 is 200 ms, 300 ms or 500 ms, the optimum value of local lag 1 is within the range from about 125 ms to around 200 ms, from about 225 ms to around 300 ms, or from about 425 ms to around 500 ms, respectively.

From Figs. 9 and 10, we notice that the contribution of the fairness is larger than that of the operability to the comprehensive quality. From Fig. 11, we can also clarify how local lags should be set to maintain the comprehensive quality as high as possible at both terminals. Local lags 1 and 2 should be set to the same values of delays 1 and 2, respectively, if the difference in network delays between the terminals is smaller than or equal to about 50 ms. When the difference is larger than about 50 ms, for simplicity, the local lags can be set to the larger value of network delays between the terminals as in the adaptive Δ-causality control.

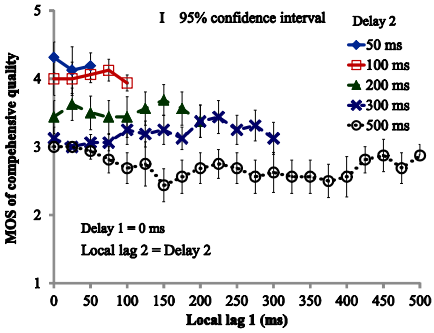


Fig. 9 MOS of comprehensive quality at terminal 1.

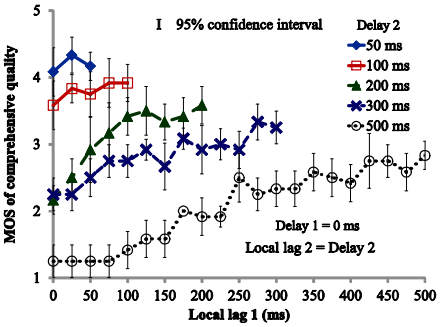


Fig. 10 MOS of comprehensive quality at terminal 2.

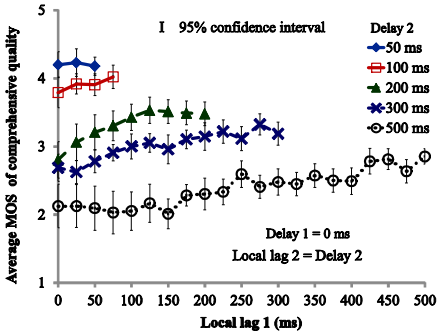


Fig. 11 Average MOS of comprehensive quality for two terminals.

5 Objective Assessment Results

We show the average numbers of burst balloons at terminals 1 and 2 in Figs. 12 and 13, respectively. The average difference in the number of burst balloons is also shown in Fig. 14. The 95% confidence intervals are also plotted in the figures.

In Fig. 12, we see that the average number of burst balloons at terminal 1 becomes smaller as local lag 1 increases for each value of delay 2. In Fig. 13, we find that the average number of burst balloons at terminal 2 hardly depends on local lag 1. By comparing Figs. 12 and 13 with Figs. 3 and 4, respectively, we can say that the MOS value of operability decreases as the average number of burst balloons becomes smaller for each value of delay 2. From Fig. 14, we notice that the average difference in the number of burst balloons becomes smaller as local lag 1 increases for each value of delay 2. This is because the local lag difference between the two terminals becomes smaller as local lag 1 increases. By comparing Fig. 14 to Fig. 5 or 6, we see that the tendencies of the curves are inverses of each other; that is, the highest MOS value can be obtained when the average difference in the number of burst balloons is the smallest for each value of delay 2.

5.1 Relations between Objective and Subjective Results

As described in Section 4.2, by comparing the objective results to subjective results, we found that they are related to each other. We should clarify how they are related to each other in detail, and examine whether the subjective results can be estimated from objective results or not. To investigate the relationship between the average number of burst balloons (or the local lag) and the MOS value of operability, we carried out the regression analysis [18]. As a result, we obtained estimated equations shown in Table 1 for both terminals of the two cases. In Table 1, O_{MOS} denotes the estimated MOS value of operability, N_{burst} is the average number of burst balloons, Δ is the local lag, and R^2 is the contribution rate adjusted for degrees of freedom [18], which shows goodness of fit with the estimated equation.

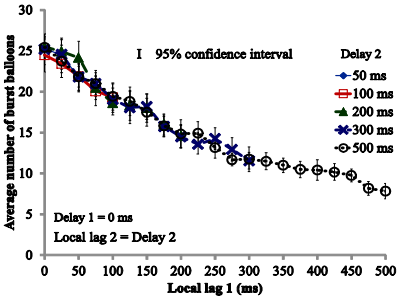


Fig. 12 Average number of burst balloons at terminal 1.

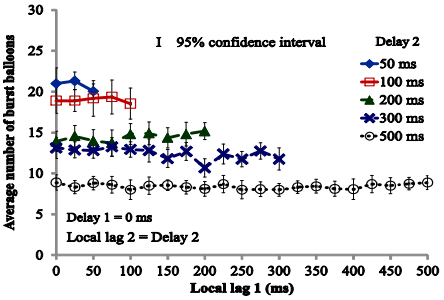


Fig. 13 Average number of burst balloons at terminal 2.

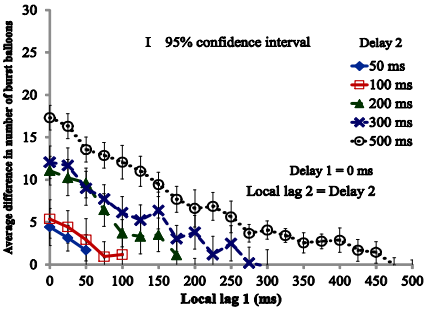


Fig. 14 Average difference in number of burst balloons between terminals.

Since the contribution rates are very high in Table 1, we can say that the MOS value of operability can be estimated with a high degree of accuracy from the average number of burst balloons or the local lag.

We also carried out regression analysis to investigate the relationship between the MOS value of fairness and the absolute value of average difference in the number of burst balloons, and that between the MOS value of fairness and the absolute value of local lag difference. As a result, we obtained equations shown in Table 2 for both terminals in the two cases. In Table 2, F_{MOS} is the estimated MOS value of fairness, $D_{N_{\mathrm{burst}}}$ is the average difference in the number of burst balloons, and D_{Δ} is the local lag difference. From Table 2, we find that the contribution rates are high. Thus, we can say that the MOS value of fairness can be estimated with a high degree of accuracy from the absolute value of average difference in the number of burst balloons or the absolute value of the local lag difference. Furthermore, according to high contribution rates shown in Table 3, where C_{MOS} is the estimated MOS value of comprehensive quality, we can say that we can estimate the MOS value of comprehensive quality with a high degree of accuracy from the average number of burst balloons and the average difference in the number of burst balloons, or from the lcoal lag and the local lag difference. In addition, we examined the relationship between the average MOS values of two

terminals and the objective results. As a result, we obtained the similar results as in Tables 1, 2 and 3.

6 Conclusions

In this paper, we investigated the trade-off relationship between the operability and fairness in a networked balloon bursting game by carrying out subjective and objective QoE assessments. As a result, we confirmed that there exists a trade-off relationship between the operability and fairness. We also clarified how local lags should be set at each terminal as follows. Local lag at each terminal can be set to the same value of network delay from the local terminal to the other terminal when the absolute difference in network delay between the terminals is smaller than or equal to about 50 ms. When the difference is larger than about 50 ms, we can set the local lags to the larger value of network delay between the terminals.

We further examined the relationships between subjective and objective assessment results. We found that the MOS value of operability can be estimated from the average number of burst balloons or local lag with a high degree of accuracy. We illustrated that the MOS value of fairness can roughly be estimated from the average

Table 1 Estimated equations for MOS of operability.

Equation	R^2
$O_{\mathrm{MOS}} = 4.347N_{\mathrm{burst}}+2.428$	0.968
$O_{\mathrm{MOS}} = -137.230\Delta+643.900$	0.968

Table 2 Estimated equations for MOS of fairness.

Equation	R^2
$F_{\mathrm{MOS}} = -4.311\left\lvert D_{N_{\mathrm{burst}}}\right\rvert + 19.226$	0.889
$F_{\mathrm{MOS}} = -135.980\lvert D_{\Delta}\rvert + 597.310$	0.936

Table 3 Estimated equations for MOS of comprehensive quality.

Equation	R^2
$C_{\mathrm{MOS}} = 0.112N_{\mathrm{burst}} - 0.107\left\lvert D_{N_{\mathrm{burst}}}\right\rvert + 1.891$	0.931
$C_{\mathrm{MOS}} = -0.003\Delta - 0.003\lvert D_{\Delta}\rvert + 4.196$	0.928

difference in number of burst balloons or the absolute value of local lag difference. Moreover, we noted that the MOS value of comprehensive quality can be estimated from the average number of burst balloons and the absolute value of average difference in number of burst balloons, or from the local lag and the absloute value of local lag differnce to a large extent.

As our future work, we will carry out QoE assessments to confirm the trade-off relationship in other networked real-time games. We also need to confirm whether the MOS value of comprehensive quality can be kept high or not at both terminals by setting local lags according to the difference in network delay between the terminals as described in Section 4.1. Furthermore, it is important to carry out the confirmation in real-time games over the Internet.

References

1. Morris, D., Joshi, N., Salisbury, K.: Haptic battle pong: high-degree-of-freedom haptics in a multiplayer gaming environment. In: Proc. Experimental Gameplay Workshop, Game Developers Conference, March 2004
2. Andrews, S., Mora, J., Lang, J., Lee, W.S.: HaptiCast: a physically-based 3D game with haptic feedback. In: Proc. Future Play Conference, October 2006
3. Ishibashi, Y., Kaneoka, H.: Fairness among game players in networked haptic environments: Influence of network latency. In: Proc. IEEE International Conference on Multimedia and Expo (ICME), July 2005
4. Ishibashi, Y., Hoshino, S., Zeng, Q., Fukushima, N., Sugawara, S.: QoE assessment of fairness in networked game with olfaction: Influence of time it takes for smell to reach player. Springer's Multimedia Systems Journal (MMSJ), Special Issue on Network and Systems Support for Games **20**(5), 621–631 (2014)
5. Mauve, M., Vogel, J., Effelsberg, W.: Local lag and timewrap: Providing consistency for replicated continuous applications. IEEE Trans. on Multimedia **6**(1), 47–57 (2004)
6. Sithu, M., Huang, P., Ishibashi, Y., Fukushima, N.: Influence of network delay on QoE for soft objects in networked haptic virtual environment. In: IEICE Technical Report, CQ2014–82, November 2014
7. Sithu, M., Ishibashi, Y., Huang, P., Fukushima, N.: QoE assessment of fairness between players for balloon bursting game in networked virtual environment with haptic sense. IEICE Technical Report, MVE2014–52, January 2015
8. Brun, J., Safaei, F., Boustead, P.: Managing latency and fairness in networked games. Communications of ACM **49**(11), 46–51 (2006)
9. Ishibashi, Y., Hashimoto, Y., Ikedo, T., Sugawara, S.: Adaptive Δ-causality control with adaptive dead-reckoning in networked games. In: Proc. the 13th Annual Workshop on Network and Systems Support for Games (NetGames), pp. 75–80, September 2007
10. ITU-T Rec. P. 10/G. 100 Amendment 1, New appendix I – Definition of quality of experience (QoE). International Telecommunication Union, January 2007
11. Fujimoto, M., Ishibashi, Y.: The effect of stereoscopic viewing of a virtual space on a networked game using haptic media. In: Proc. ACM SIGCHI International Conference on Advances in Computer Entertainment Technology (ACE), pp. 317–320, June 2004
12. http://geomagic.com/en/products/phantom-omni/overview
13. SensAble Technologies, Inc., Openhaptics toolkit programmer's guide, version 3.0 (2009)
14. Carson, M., Santay, D.: NIST Net–A Linux-based network emulation tool. ACM SIGCOMM **33**(3), 111–126 (2003)
15. Ishibashi, Y., Tasaka, S., Hasegawa, T.: The Virtual-Time Rendering algorithm for haptic media synchronization in networked virtual environments. In: Proc. the 16th International Workshop on Communications Quality and Reliability (CQR), pp. 213–217, May 2002
16. Sithu, M., Ishibashi, Y., Fukushima, N.: Effects of dynamic local lag control on sound synchronization and interactivity in joint musical performance. ITE Trans. Media Technology and Applications, Special Section on Multimedia Transmission System and Services **2**(4), 299–309 (2014)
17. ITU-R BT. 500-12, Methodology for the subjective assessment of the quality of television pictures. International Telecommunication Union, September 2009
18. Chatterjee, S., Hadi, A.S., Price, B.: Regression analysis by example. John Wiley & Sons, Hoboken (1999)

The Effect of Spatiotemporal Tradeoff of Picture Patterns on QoE in Multi-View Video and Audio IP Transmission

Toshiro Nunome and Yusuke Tsuya

Abstract In this paper, we study QoE (Quality of Experience) of Multi-View Video and Audio (MVV-A) transmission over IP networks. This paper assesses the effect of the tradeoff relationship between improvement of image quality and degradation of viewpoint change response owing to the picture patterns. When the length of GOP (Group of Picture) is short, the viewpoint change response is quick, but the image quality is not good. On the other hand, in the long GOP, the image quality is good owing to high coding efficiency, while the viewpoint change response is slow because the new viewpoint cannot be shown until receiving the next I picture. We employ two contents and assess QoE multidimensionally by a subjective experiment.

1 Introduction

For giving higher presence to the users, *MVV (Multi-View Video)*, in which the users can change the viewpoint, has been studied [1]. The ultimate goal of the network services is to provide high *QoE (Quality of Experience)*, which represents the overall acceptability of an application or service, as perceived subjectively by the end-users [2].

In [3] and [4], *MVV-A (MVV and Audio)*, which is MVV accompanied by audio, is transmitted over the IP network, and QoE assessment has been conducted. However, the papers only consider I pictures only for the picture pattern of the video stream. When we employ P pictures, i.e., utilizing inter-frame prediction for encoding, even

T. Nunome(✉)
Department of Computer Science and Engineering, Graduate School of Engineering, Nagoya, Japan
e-mail: nunome@nitech.ac.jp

Y. Tsuya
Department of Computer Science, Faculty of Engineering, Nagoya Institute of Technology, Nagoya 466-8555, Japan

T.T. Zin et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 388,
DOI: 10.1007/978-3-319-23207-2_14

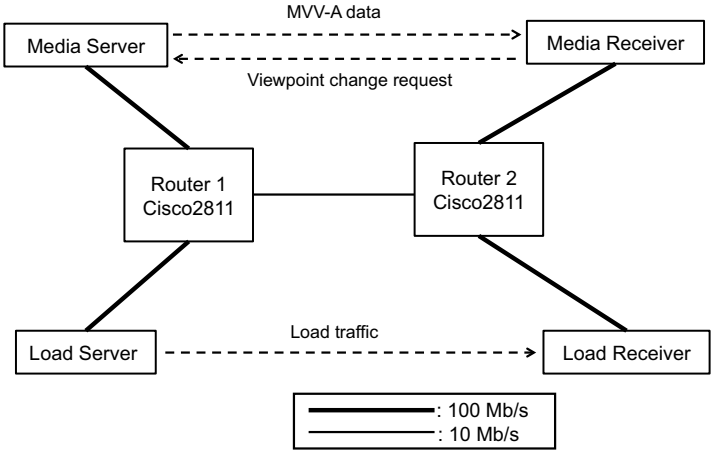


Fig. 1 Experimental system

if the users change the viewpoint, the new viewpoint cannot be shown until receiving the next I picture. It degrades viewpoint change response. On the other hand, the inter-frame prediction improves coding efficiency; it can provide high image quality.

Thus, in this paper, we assess the effect of the tradeoff relationship between improvement of image quality and degradation of viewpoint change response owing to the picture patterns. We assess QoE multidimensionally by a experiment with two contents.

The remainder of this paper is organized as follows. Section 2 describes the experimental method. Section 3 explains the QoE assessment method. Section 4 presents experimental results. Section 5 concludes this paper.

2 Experimental Method

Fig. 1 shows the configuration of the experimental system. Media Server is the server of MVV-A, and Media Receiver is the client of MVV-A. Load Server is the server of the load traffic, and Load Receiver is the client of the load traffic. Both Router 1 and Router 2 are Cisco 2811. Two routers are connected by a full duplex Ethernet line of 10 Mb/s. All the other links are 100 Mb/s Ethernet.

The server captures the video of each camera. At the same time, the audio is captured by the microphone. The server sends the audio and video of a selected viewpoint to the client as two separate UDP packet flows. The client receives these packets and outputs the audio and video decoded from them. The client can choose one viewpoint from the four cameras by sending a request with a UDP packet.

The specifications of audio and video are shown in Table 1. We refer to the transmission unit at the application-level as an MU (Media Unit). A video MU is a video frame and an audio MU is 320 audio samples. Each MU is transmitted as

Table 1 Specifications of audio and video.

	audio	video
coding method	G.711 μ-law	H.264
picture size [pixels]	-	704 $\times$ 480
picture pattern	-	I, IPPPP, I+14P's
encoding bit rate [kb/s]	64	1000
average MU rate [MU/s]	25	30
duration [s]	20	

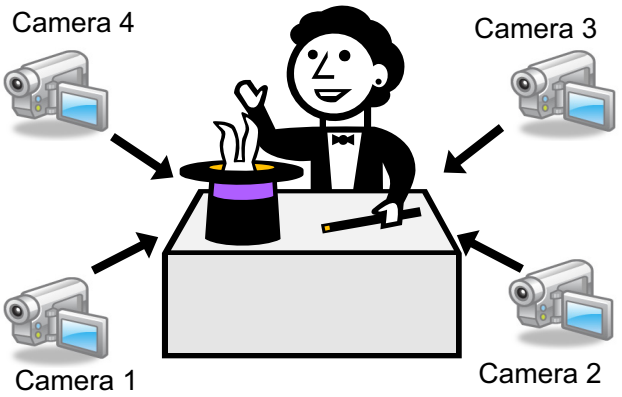


Fig. 2 Camera arrangement (magic)

a UDP packet. We employ frame skipping as the output method of video. That is, when some packets consisting of an MU is lost, output of the MU is skipped.

Load Server generates UDP packets including 1472 bytes payload with exponentially distributed interval and sends them to corresponding Load Receiver. We assume two values of the average amount of UDP load traffic: 5.9 Mb/s and 8.4 Mb/s. We consider that the average load traffic 5.9 Mb/s and 8.4 Mb/s are lightly loaded and heavily loaded, respectively.

In this paper, playout buffering control is used for absorbing delay jitter in Media Receiver. In the MVV-A system, playout buffering control brings trade-off between the viewpoint change response and output quality [3]. In order to investigate the effect of the playout buffering time, we employ three values: 70, 150, and 300 ms.

In the experiment, we employ two contents. One is a table magic with cards (namely, magic). The other is a toy train running on plastic rails (namely, train). Figs. 2 and 3 show the camera arrangements in magic and train, respectively.

We employed 15 male students in their twenties as assessors. We have totally 36 stimuli to be evaluated because of the two contents, the three picture patterns, the two patterns of load traffic, and the three types of playout buffering time. The total time for the experiment to an assessor is about 40 minutes.

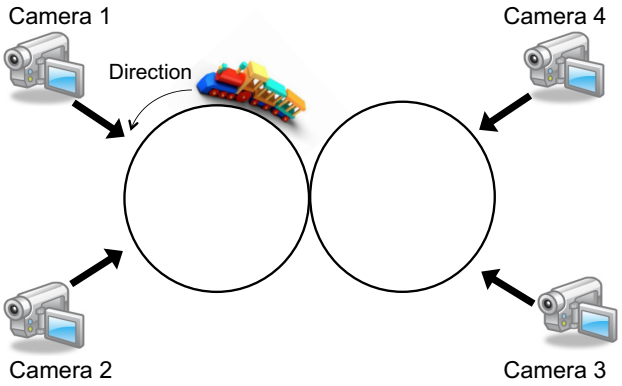


Fig. 3 Camera arrangement (train)

3 QoE Assessment Method

In the experiment, we perform multidimensional QoE assessment with 10 pairs of polar terms. The pairs in the subjective experiment are shown in Table 2 and are classified into six categories; there are three pairs for video, a pair for audio, a pair for synchronization, two pairs for response, two pairs for psychology, and a pair for overall satisfaction.

Table 2 Adjective pairs

category	adjective pair
video	The video is rough - smooth
	The video is blurred - sharp
	The video is hard to grasp - easy
audio	The audio is artificial - natural
synchronization	The audio and video are out of synchronization - in synchronization
response	The viewpoint change response is slow - fast
	The viewpoint change response is unstable - steady
psychology	I am irritated - not irritated
	I feel impatient - relaxed
overall	Bad - Excellent

Note that the experiment was performed with the Japanese language. This paper has translated the used Japanese terms into English. Therefore, the meanings of adjectives or verbs written in English here may slightly differ from those of Japanese ones.

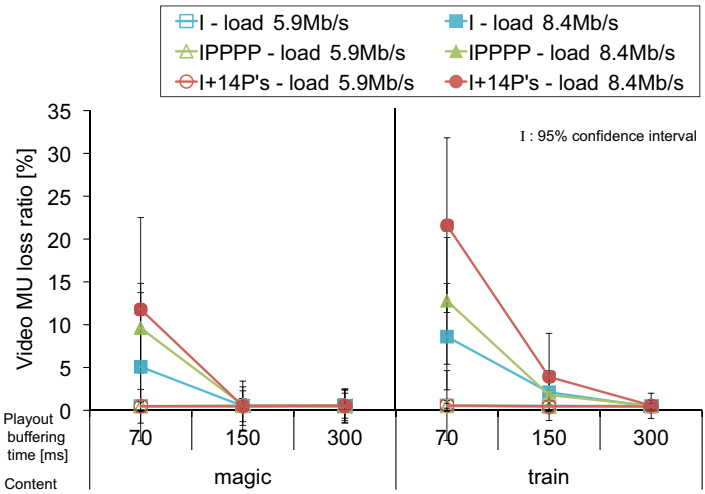


Fig. 4 Video MU loss ratio

For each pair of polar terms, the subject gives a score to the stimulus by the *rating scale method* [5] with five grades. The best grade (score 5) represents the positive adjective (the right-hand side one in each pair in Table 2), while the worst grade (score 1) means the negative adjective. The middle grade (score 3) is neutral. Finally, we calculate the mean opinion score (MOS), which is average of the rating scale scores for all the users.

4 Experimental Result

4.1 Application-Level QoS

Figure 4 shows the MU loss ratio of video. It is the ratio of the number of MUs not output at the recipient to the number of MUs transmitted by the sender. The abscissa means the combination of the playout buffering time and the content. Each plot shows the result for the combination of the picture pattern and the average load traffic.

We see in Fig. 4 that under lightly loaded condition, the MU loss merely occurs. On the other hand, under heavily loaded condition, we notice that the MU loss occurs for the playout buffering time 70 ms; the MU loss ratio decreases as the playout buffering time increases. This is because the small buffering time cannot absorb network delay jitter under the condition.

We also find in Fig. 4 that as the GOP length becomes long, the MU loss ratio increases under heavily loaded condition. This is because the video output is skipped until the next I picture when an MU cannot be received. In addition, the MU loss

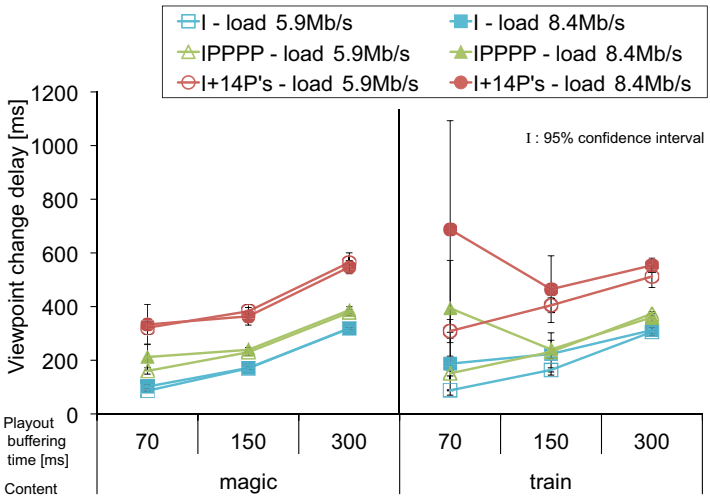


Fig. 5 Viewpoint change delay

ratio of train is larger than that of magic. This is due to the characteristics of contents; the toy train moves largely in train, and the movement is small in magic.

Figure 5 depicts the average viewpoint change delay. It is defined as the time in seconds from the moment the user inputs a request for viewpoint change by the user interface until the instant a new viewpoint is output at the client.

In Fig. 5, the viewpoint change delay is approximately the summation of the round trip delay between the server and the client and the playout buffering time. That is, the playout buffering time degrades the response of the viewpoint change.

On the other hand, under heavily loaded condition with small buffering time, the viewpoint change delay is large. This is because one or more MUs of a new viewpoint can be discarded owing to their delayed arrival, and then the video freezes until the new viewpoint is displayed. In this situation, train has larger viewpoint change delay than magic. This is also the effect of the characteristics of contents.

4.2 QoE

In this section, we picked up the results of two adjective pairs. Figure 6 shows the MOS for "The viewpoint change response is slow - fast". Figure 7 depicts the MOS for "Bad - Excellent".

In Fig. 6, we find that the viewpoint change response for picture pattern I is faster than the other picture patterns especially in short buffering time. This is because with the picture patterns IPPPP and I+14P's, the client needs to wait until the next I picture when the viewpoint change request occurs.

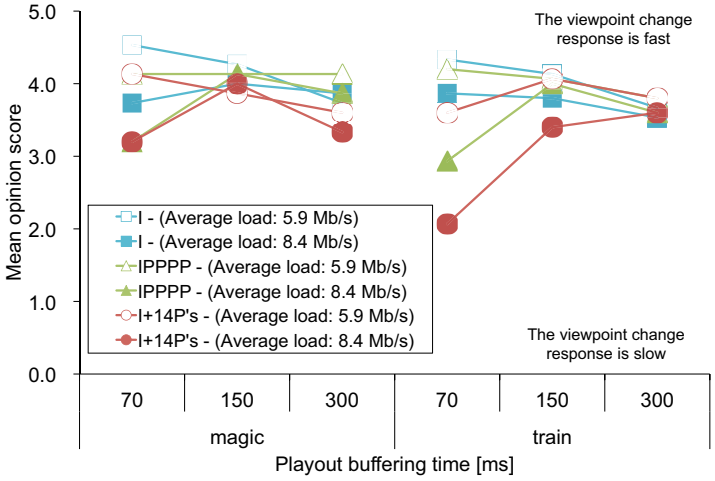


Fig. 6 Viewpoint change response

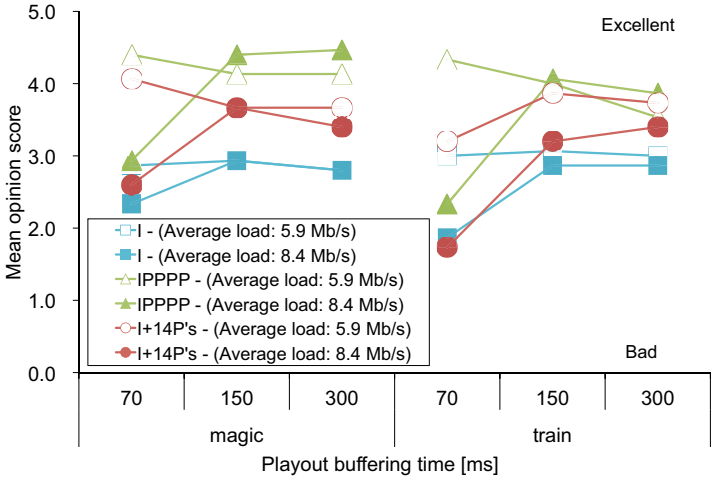


Fig. 7 Overall satisfaction

We also see in Fig. 6 that the viewpoint change response for train is lower than that in magic. This is because train has larger action than magic and then requires quick response for viewpoint change.

In Fig. 7, we can observe that the overall satisfaction for picture pattern IPPPP is the best among the three picture patterns considered in this paper. This is because the picture pattern has a good balance between the viewpoint change response and picture quality.

We notice in Fig. 7 that when the playout buffering time is set to 70 ms under heavily loaded condition, the overall satisfaction for all the picture patterns degrades. This is because the short buffering time cannot absorb network delay jitter.

In addition, we find in Fig. 7 that regardless of the average load traffic, the picture pattern I has smaller MOS values than the other two picture patterns. This is because the picture pattern has low coding efficiency and then cannot provide high picture quality with the encoding bit rate in this paper, i.e., 1000 kb/s.

As for the effect of the contents, we see that the difference of MOS values between train and magic is small for the playout buffering time 150 ms and 300 ms. This implies that the spatiotemporal tradeoff differs by the contents; train has larger weight of viewpoint change response than magic. Thus, the overall satisfaction of IPPPP and I+14P's in train becomes smaller than that in magic.

5 Conclusions

In this paper, we evaluated the effect of the tradeoff relationship between improvement of image quality and degradation of viewpoint change response owing to the picture patterns. As a result, we noticed that the effect of spatiotemporal quality tradeoff on QoE differs from the contents.

In future work, we need to assess QoE in other picture patterns. In addition, we will devise appropriate setting methods of picture patterns.

Acknowledgments We thank Professor Emeritus Shuji Tasaka for his valuable discussion.

References

1. Ahmad, I.: Multiview video: get ready for next-generation television. In: Proc. IEEE Distributed Systems Online, vol. 8, no. 3, art. no. 0703–o3006, March 2007
2. ITU-T Rec. P.10/G.100, Amendment 2. New definitions for inclusion in Recommendation ITU-T P.10/G.100, July 2008
3. Rodriguez, E.J., Nunome, T., Tasaka, S.: QoE assessment of multi-view video and audio IP transmission. IEICE Trans. on Commun. **E93–93**(6), 1373–1383 (2010)
4. Rodriguez, E.J., Nunome, T., Tasaka, S.: Multidimensional QoE assessment of multi-view video and audio (MVV-A) IP transmission: the effect of user interfaces and contents. In: Proc. IEEE WAINA 2012, pp. 91–98, March 2012
5. Guilford, J.P.: Psychometric methods. McGraw-Hill, N.Y. (1954)

Anomalous Behavior Detection in Mobile Network

Mon Mon Ko and Mie Mie Su Thwin

Abstract New security threats emerge against mobile devices as the devices' computing power and storage capabilities evolve. Preventive mechanisms like authentication, encryption alone are not sufficient to provide adequate security for a system. In this work, we propose User Group Partition Algorithm and Behavior Pattern Matching Algorithm to extract anomalous calls from mobile call detail records effectively. The system accepts the proper input of normal mobile phone call detail records as training dataset and fraud mobile phone call detail records as testing dataset. Two main processes are included in this system: grouping mobile phone calls in training dataset according to similar phone call patterns and matching the new input mobile phone call detail records with grouped mobile phone call patterns to examine the input mobile phone call detail record is normal or not. If the system detects the anomalous mobile phone behavior, the system warns the user that the suspicious mobile phone call is detected and asks the user which action will be taken.

1 Introduction

Mobile devices have evolved and experienced a great success over the last few years. Such devices are capable of performing sophisticated tasks and communicate through various wireless interfaces [1]. However, along with their popularity, mobile devices face an everyday growing number of security threats. This is despite the variety of peripheral protection mechanisms proposed in the literature in recent years. Without doubt, authentication and access control methods can be

M.M. Ko(✉)
University of Computer Studies, Mandalay, Myanmar
e-mail: nannlaypyaenu@gmail.com

M.M.S. Thwin
Myanmar Computer Emergency Response Team,
Ministry of Science and Technology, Yangon, Myanmar

T.T. Zin et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 388,
DOI: 10.1007/978-3-319-23207-2_15

used in many cases, but alone, they are not sufficient to offer integral protection against intrusions [2]. Overall, with the increasing risk of mobile malware, the theft or loss of mobile devices and the physical vulnerability, i.e. rewiring a circuit on the chip or using probing pins to monitor data flows to retrieve private keys or find flaws in the hardware components, designing a highly secure mobile device is still a very challenging task [3].

While more than four billion people enjoy their mobile devices using 2G/3G mobile networks, Kaspersky Lab has very recently identified 39 new mobile malware families (SMS trojans, iPhone malware, Android spyware) with 143 modifications [4]. According to a Scan Safe report malware volumes grew 300% in 2008, and it is noted that most of the legitimate web pages crawling on the Internet are not trustworthy or infected by different kinds of viruses [5]. Moreover, according to the UK Home Office, 69% of robberies include a mobile device. As a result, a need for more intelligent and sophisticated security controls such as Intrusion Detection Systems (IDSs) for mobile devices is necessary[6]. In general, there are two basic approaches in IDS to detect an intrusion: a) misuse based (also called signature-based), and b) anomaly based (also called behavior-based). Although misuse based IDS can immediately be employed to monitor the mobile environment, only an anomaly-based IDS is able to detect new, unforeseen vulnerabilities and variants of known attacks [7]. Anomaly-based intrusion detection profiles normal behavior and attempts to identify patterns of user activities that deviate from a predefined or dynamically updated profile [8, 9]. Whilst much research has been devoted to IDS, in the context of anomaly detection, the exploration of what is defined as "normal" has been limited and several important problems remain unsolved [10, 11].

In this paper we concentrate on anomaly-based IDS for mobile network. We use a data set generated from a database of real world mobile communication network information. The database provides the following information for each transaction (use of a service by a customer): the initiation time, the duration (in minutes), and the type of the service. From the database, we generated a data set with 13,280 examples.

1.1 Motivation

The motivation behind this approach is to seek proper as opposed to normal behavior. It also to overcome the drawbacks of existing approaches such as specification increasing, high false alarm rate, and to provide the solve options for users.

1.2 Purposes of the System

The purposes of this System are:

- To study anomaly detection in mobile network elaborately.
- To propose an effective detection system using GP algorithm and BPM algorithm and summarizing mobile user behavior.
- To eliminate the drawbacks of previous works in anomaly detection in mobile network.

- To develop the applicable anomaly detection system of mobile network effectively.
- To help for getting high accuracy in detection of anomalies in call detail records.
- To highlight the importance of anomaly detection process for mobile telecommunication network.
- To observe the usefulness of mobile user behavior patterns.
- To pinpoint the impacts of CDR features for mobile user behavior monitoring.

2 Literature Review

This chapter presents some knowledge and overview of approaches in the literature concerned with anomalous behavior detection in mobile phone system.

The work in proposed a prototype of a tool, based on a supervised Artificial Neural Network (ANN), to detect anomalous behavior on mobile communications, such as service fraud and Subscriber Identity Module (SIM) card cloning [12]. The authors, based on their prototype, report accuracy of a 92.50% detection of fraudulent users and a 92.5% correct classification of legitimate users. The work in proposed the Bayes Decision Rule (BDR) towards the generation of mobility user profiles within the Global System for Mobile Communications (GSM) network [13]. By utilizing their method the authors managed to achieve a TPR of 83.50%. One problem with this approach is the privacy of the end-user's usage log files, which are exposed to the telecom carriers in order to detect mistrusted users, as explained in [14].

Hollmen has proposed fraud detection techniques in mobile networks by means of user profiling and classification [15]. Specifically, the author used ANN and probabilistic models to detect anomalous usage and achieved a TPR of 69%. However, the presented method for fraud detection is based on an available large database with billions of records. As a result, this method can be seen only as a specific user profiling problem in fraud detection. The authors in used ANN to form short and long-term statistical behavior profiles for GSM and Universal Mobile Telecommunication Systems (UMTS) networks [16]. They define two time spans over the call data records, i.e. a shorter sequence or Current Behavior Profile (CBP) and a longer one or Behavior Profile History (BPH).

Also, the authors proposed an on-line anomaly detection algorithm, based on Markov Model, where the key distinguishing characteristic is the use of sequences of network cell IDs traversed by a user [17]. Recently, the authors presented a tested for experimenting with anomaly detection algorithms and demonstrated its properties using two unsupervised anomaly detection methods, i.e. Self-Organizing Map (SOM) and clustering [18]. They conclude that both methods are suitable for network monitoring.

The work presented a behavioral detection framework for malware targeting mobile devices [19, 20]. Particularly, the framework generates a malicious behavior signature database by extracting the key behavior signatures from the mobile malware.

3 Research Methodology

The objective of this Research is to develop an anomalous behavior detection system for mobile network, should be able to effectively detect the anomalous call and messages in mobile network communication. In order to accomplish this objective, a study of existing approaches to anomalous behavior detection in mobile network is conducted. The problem of anomalous behavior detection in mobile network has been investigated by researchers and many kinds of anomalous behavior detection methods have been proposed. In general, there are two Approaches which can be observed.

The first Approach is Anomaly-based Detection. Most of the Intrusion Detection Methods use the Anomaly-based intrusion detection. Actually Anomaly detection relies on models of the intended behavior of users and applications and interprets deviations from this 'normal' behavior as evidence of malicious activity. The second Approach, Specification-based intrusion detection defines the precise expected behavior of the system for specific events like messaging and calling.

Actually anomalous behavior detection for mobile phone network is needed to reduce the drawbacks of previous works in anomaly detection in mobile phone network. And a relatively lightweight and fast method is also needed to do the grouping and matching mobile phone call patterns. So in this System, the mobile phone call detail records are used as inputs to effectively detect the anomalous mobile phone call behavior.

3.1 Proposed Approach

We have two key ideas for our proposed approach: a call detail record user assessed and how the anomalous call record has been detected. Based on these ideas, call detail record; there are three main steps for detecting the anomalous call record behavior. They are 1) grouping the normal mobile phone users according to their mobile phone usage, 2) matching the abnormal mobile phone user with grouped mobile user, 3) determining the new mobile phone transaction is normal or not.

First of all, we apply data preprocessing step to eliminate the mobile phone transaction which duration is less than or equal to 0 and if the CDR is not complete CDR.

Later, we apply the user group partitioning algorithm to group the mobile phone users with similar mobile phone usage.

```
Alogorithm    : CDR_Partition
Input         : Interval, CDR
Output        : Group k
Begin
      Group_list = Null
      For I= 0 to i<CDR.Count
```

```
    {
        If group_list = Null
        {
         Group_list = Group1;
         Int_In_Count_Min = CDR(i).Int_In_Count – Interval
         Int_In_Count_Max = CDR(i).Int_In_Count + Interval
         Int_In_Duration_Min = CDR(i).Int_In_Duration – Interval
         Int_In_Duration_Max = CDR(i).Int_In_Duration + Interval
         Int_Out_Count_Min = CDR(i).Int_Out_Count – Interval
         Int_Out_Count_Max = CDR(i).Int_Out_Count + Interval
         Int_Out_Duration_Min = CDR(i).Int_Out_Duration – Interval
         Int_Out_Duration_Max = CDR(i).Int_Out_Duration + Interval
         If Int.Data <0 then Int.Data = 1;
         Group1.Data = Int.Data
    }
    Else if group_list > 0
    {
    If check = false
    {
    Group_list = Group_list.Count + 1;
        Int_In_Count_Min = CDR(i).Int_In_Count – Interval
        Int_In_Count_Max = CDR(i).Int_In_Count + Interval
        Int_In_Duration_Min = CDR(i).Int_In_Duration – Interval
        Int_In_Duration_Max = CDR(i).Int_In_Duration + Interval
        Int_Out_Count_Min = CDR(i).Int_Out_Count – Interval
        Int_Out_Count_Max = CDR(i).Int_Out_Count + Interval
        Int_Out_Duration_Min = CDR(i).Int_Out_Duration – Interval
        Int_Out_Duration_Max = CDR(i).Int_Out_Duration + Interval
                If Int.Data <0 then Int.Data = 1;
                Group1.Data = Int.Data
            }
        }
    }
End
```

Then, we apply the behavior pattern matching algorithm to detect the anomalous mobile phone calls.

```
Algorithm: Matching_Call_Behavior
Input      : Group_Data, Test_Data
Output     : Maximum Similarity
Begin
    Similarity = 0;
    X = 0.1;
    For i=0: i< Group_Data.Count; i++
        For each index (j) of Group_Data and Test_Data
        If index(j) holds odd number
```

```
{
    If Group_Data(j) == Test_Data(j) then
    Similarity = Similarity +1;
  If index(j) holds even number
   If Test_Data(j) <= (1-X) * Group_Data(j) and
        Test_Data(j) >= (1+X) * Group_Data(j) then
    Similarity = Similarity+1;
For k = 0; k< Test_Data(Count); k++
  If Test_Data(k).Similarity <= 4 then
  Display "Abnormal";
  If Test_Data(k).Similarity >4 then
  Display "Normal";
```

4 Evaluation Analysis of System

Evaluation performed over 100000 training datasets with 1000-10000 testing. The system is implemented by taking C# programming language over Visual Studio 2008 and Microsoft SQL 2008 Database. The experimenting environment is Laptop of Window 7, 64 bit operating system Intel(R) core(TM) i5 with internal memory 4G RAM.

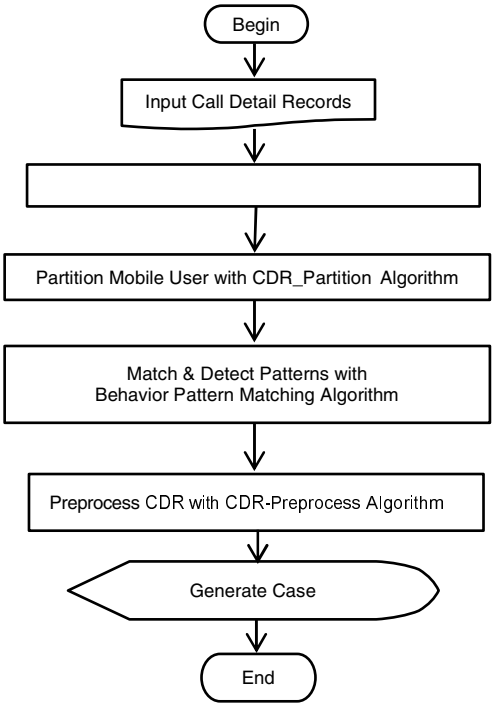


Fig. 1 Flow Diagram of the System.

The evaluation is performed over the 10000 dataset. To train the dataset, there is a need to define the suitable value for the increase or decrease count for interval and duration. By changing the different interval 1 to 25, we found that the smaller the interval, the larger the interval. So, we define the suitable interval is 10 with the 53 group for testing. The next section of testing evaluation, we used the above interval.

The detailed analysis is carried out over the calling dataset100330. Evaluation performed over 100000 training datasets with 1000-10000 testing. To train the data, we need to define the interval and the value of X. It is performed with the 10 interval and the X is 0.1-0.5. The evaluation result is indicated by the following Tables 1 with the result of Little Suspicious, Suspicious, Neglectable, Serious and Most Serious of 1000 testing dataset.

Table 1 Comparison of X over 1000 testing

X	Little Suspicious	Suspicious	Neglectable	Serious	Most Serious
0.1	38.3	16.8	23.9	15.4	5.6
0.2	38.8	17	24.3	14.6	5.3
0.3	39	16.9	24.6	14.2	5.3
0.4	39.2	17	24.8	14.1	5.1
0.5	39	17.1	24.9	14	5

Figure 2 illustrate the result of Table 1. In figure, red color describes most serious, orange describes serious, green describes neglect, yellow describes suspicious and blue describes little suspicious. The x axis describes the value of X (0.1-0.5) and the y axis describes the percentage (%).

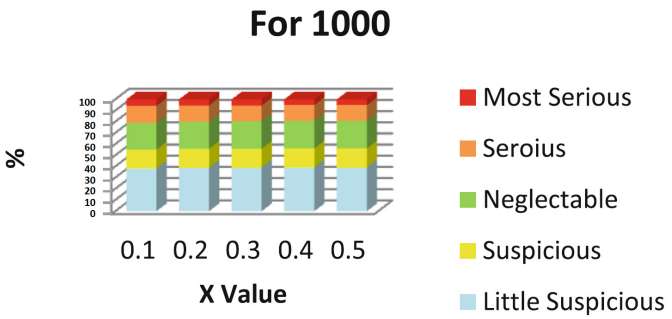


Fig. 2 Evaluation of X values 0.1- 0.5 over 1000 dataset

5 Conclusions

Call Detail Record Data contain abnormal call record, anomalies that could negatively affect the mobile phone users and network carrier. Detecting anomalous

behavior call record is an interesting problem. By using the Proposed Group Partition Algorithm and Behavior Pattern Matching Algorithm, the sudden changed call behavior can be detected effectively. Detecting the anomalous call behavior can help for many research areas of mobile phone network.

References

1. Artail, A.H., Raydan, M.: Device-aware desktop web page transformation for rendering on handhelds. Personal and Ubiquitous Computing **9**(6), 368–380 (2005). doi:10.1007/s00779-005-0348-5
2. Chow, G.W., Jones, A.: A framework for anomalydetection in OKL4-linux based smartphones. In: Proceedings of the 6th Australian Information Security Management Conference (2008)
3. Sun, B., Chen, Z., Wang, R., Yu, F., Leung, V.C.M.: Towards adaptive anomaly detection in cellular mobile networks. In: Proceedings of the IEEE Consumer Communications and Networking Conference (CCNC 2006), vol. 2, pp. 666–670 (2006)
4. Sun, B., Xiao, Y., Wu, K.: Intrusion Detection in Cellular Mobile Networks. Book chapter in Wireless Mobile Network Security, pp. 183–210. Springer (2007). ISBN: 0387280405
5. Naumann, I., Hogben, G., Fritsch, L., Benito, R., Dean, R.: Security Issues in the Context of Authentication Using Mobile Devices (Mobile eID). European Network and information Security Agency (ENISA), January (2008)
6. GSM. World Mobile. Market Data Summary (Q22009). http://www.gsmworld.com/newsroom/market-ata/market_data_summary.htm (accessed 16 Feb. 2011)
7. Mobile World Congress. Visit Kaspersky Lab at Mobile World Congress 2009 in Barcelona. http://www.kaspersky.com/news?id=207575745 (accessed 16 Feb. 2011)
8. Landesman, M.: The World's Largest Security Analysis of Real-World Web Traffic, Annual Global Threat Report, ScanSafe STAT. http://www.scansafe.com/downloads/gtr/2009_AGTR.pdf (accessed 16 Feb. 2011)
9. Ray, B.: Home Office discusses thief-proof phones. http://www.theregister.co.uk/2007/05/25/home_office_phone_crime (accessed 16 Feb. 2011)
10. Kruegel, C., Valeur, F., Vigna, G.: Intrusion Detection and Correlation: Challenges and Solutions. Book chapter Computer security and Intrusion Detection. Springer (2005)
11. Singh, K.K.: Hybrid Profiling Strategy for Intrusion Detection, Department of Computer Science. University of British Columbia (2004)
12. Moreau, Y., Verrelst, H., Vandewalle, J.: Detection of mobile phone fraud using supervised neural networks: a first prototype. In: Gerstner, W., Hasler, M., Germond, A., Nicoud, J.-D. (eds.) ICANN 1997. LNCS, vol. 1327, pp. 1065–1070. Springer, Heidelberg (1997)
13. Buschkes, D., Kesdogan, R., Reichl, P.: How to increase security in mobile networks by anomaly detection. In: Proceedings of the Computer Security Applications Conference, Phoenix, pp. 3–12, December 1998
14. Boukerche, A., Notare, M.S.M.A.: Behavior-Based Intrusion Detection in Mobile Phone Systems. Journal Of Parallel and Distributed Computing **62**(9), 1476–1490 (2002)

15. Hollmén, J.: User profiling and classification for fraud detection in mobile communications networks, PhD Thesis, Helsinki University of Technology (2000)
16. Burge, P., Shawe-Tylor, J.: An unsupervised neural network approach profiling the behavior of mobile phone users for use in fraud detection. Journal of Parallel and Distributed Computing **61**(7), 915–925 (2001)
17. Sun, B., Yu, F., Wu, K., Leung,V.C.M.: Mobilitybased anomaly detection in cellular mobile networks. In: Proceedings of the ACM Wireless Security (WiSe 2004), Philadelphia, PA, pp. 61–69 (2004)
18. Kumpulainen, P., Htnen, K.: Anomaly Detection Algorithm Test Bench for Mobile Network Management. Tampere University of Technology (2008)
19. Bose, A., Hu, X., Shin, K.G., Park, T.: Behavioral detection of malware on mobile handsets. In: Proceedings of the 6th International Conference on Mobile Systems, Applications, and Services (MobiSys 2008), USA, June 2008
20. Liu, L., Yan, G., Zhang, X., Chen, S.: VirusMeter: preventing your cellphone from spies. In: Kirda, E., Jha, S., Balzarotti, D. (eds.) RAID 2009. LNCS, vol. 5758, pp. 244–264. Springer, Heidelberg (2009)

Detection of Web Application Attacks with Request Length Module and Regex Pattern Analysis

Ei Ei Han

Abstract Web application attack detection is one of the popular research areas during these years. Security for web application is necessary and it will be effective to study and analyze how malicious patterns occur in web server log. This system analyzes web server log file, which includes normal and malicious users' access patterns with their relevant links. This uses web server log file dataset for the detection of web application attacks. This system intends to analyze normal and attack behaviors from web server log and then classify attack types which are included in the dataset. In this system, three types of attacks are detected namely, SQL injection, XSS and directory traversal attacks. Attack analysis stage is done by request length module and regular expressions for various attack patterns.

1 Introduction

Web applications are becoming increasingly popular and complex in all sorts of environments, ranging from e-commerce applications to banking. The security of web applications has become increasingly important and a secure web environment has become a high priority for e-business communities. They are subject to all sorts of attacks. In today's times, the most critical issue for any web application is security. Web servers and web-based applications are popular attack targets. To detect web-based attacks, intrusion detection systems are configured with a number of signatures that support the detection of known attacks.

This system differentiates normal access patterns from malicious access patterns. It can detect how malicious users try to attack the web site. The system

E.E. Han(✉)
University of Computer Studies, Yangon, Myanmar
e-mail: eieihan.ucsy@gmail.com

T.T. Zin et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 388,
DOI: 10.1007/978-3-319-23207-2_16

can know which pages or links are most accessed and are tried by malicious users. It also describes successful attacked (attack gained) web pages and links. This system will be effective for the security of web application system and analysis on web server log. There are two fundamentally different attack detection methods – rule-based detection (static rules) and anomaly-based detection (dynamic rules).

Web server log analysis is a rule-based detection mode which concentrates on web attacks which are visible in default web server log files like Apache or IIS. This system combines traditional web usage mining system with security analysis.

2 Background Theory

2.1 Web Usage Mining

Web usage mining is the process of extracting useful information by analyzing web usage data from server logs. It is defined as an application of data mining techniques on the navigational traces of the users to extract knowledge about their preferences and behavior. Web usage mining involves three major phases namely, pre-processing, pattern discovery and pattern analysis. Some of the techniques used in Pattern discovery are Association rules, Classification, Clustering etc. Pattern Analysis filters out uninteresting rules or patterns found in the pattern discovery phase.

2.2 Web Application Attacks and DVWA

Malicious users try to attack a web site or web server by using various attack patterns. Web application attacks are occurred by performing Web application queries. They take the forms of well defined strings and parameters. These are recorded in the web server log file. By analyzing each record of server log file, malicious patterns can be detected. These patterns include some special and encoding characters. To classify the web based attacks, it is needed preparing for the input data like URL decoding and regular expression, etc.

One of the popular web application attack tools is DVWA. Damn Vulnerable Web App (DVWA) is a PHP/MySQL web application that is damn vulnerable. Its main goals are to be an aid for security professionals to test their skills and tools in a legal environment, help web developers better understand the processes of securing web applications. It will be used for launching web application attacks and logging them. With this tool, popular web based attacks can be created and stored in the database.

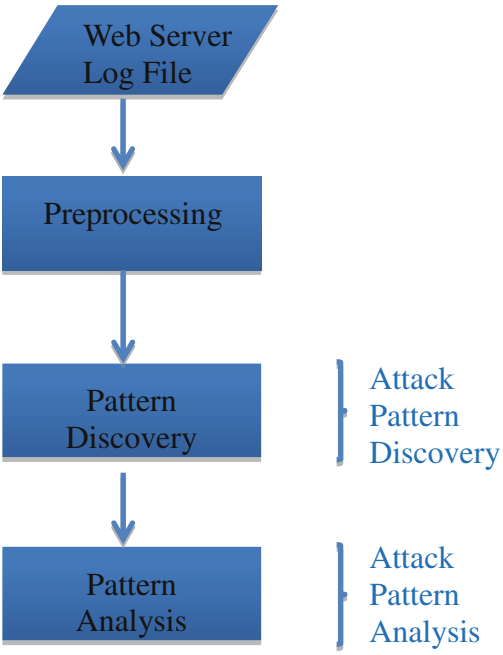


Fig. 1 Web Usage Mining Framework Combined with Intrusion Detection

Examples of web server log file by testing with DVWA are as follows:

127.0.0.1 - [23/Apr/2014 12:55:31 +0630] "GET /DVWA-1.0.8/ HTTP/1.1" 200 4618 "-" "Mozilla/5.0 (Windows NT 6.1; rv:27.0) Gecko/20100101 Firefox/27.0"
127.0.0.1 - [23/Apr/2014 12:55:35 +0630] "GET /DVWA-1.0.8/vulnerabilities/xss_r/ HTTP/1.1" 200 4456 "http://localhost/DVWA-1.0.8/" "Mozilla/5.0 (Windows NT 6.1; rv:27.0) Gecko/20100101 Firefox/27.0"
127.0.0.1 - [23/Apr/2014 12:55:50 +0630] "GET /DVWA-1.0.8/vulnerabilities/xss_r/?name=%3Cscript%3Ealert%28%2Fxss%2F%29%3C%2Fscript%3E HTTP/1.1" 200 4514 "http://localhost/DVWA-1.0.8/vulnerabilities/xss_r/" "Mozilla/5.0 (Windows NT 6.1; rv:27.0) Gecko/20100101 Firefox/27.0"
127.0.0.1 - [23/Apr/2014 12:56:20 +0630] "GET /DVWA-1.0.8/vulnerabilities/xss_r/?name=%3CScriPt%3Ealert%28%2Fxss%2F%29%3C%2FScripT%3E HTTP/1.1" 200 4514 "http://localhost/DVWA-1.0.8/vulnerabilities/xss_r/?name=%3Cscript%3Ealert%28%2Fxss%2F%29%3C%2Fscript%3E" "Mozilla/5.0 (Windows NT 6.1; rv:27.0) Gecko/20100101 Firefox/27.0"

The web server log file which contains attacks includes URL encoding characters. So, it is needed to decode these characters to get attack patterns. Percent ("%") character serves as the indicator for percent-encoded octets. It is the practice of translating unprintable characters or characters with special meaning within URLs to a representation that is unambiguous and universally accepted by web browsers and servers. When you pass information through a URL, you need to make sure it only uses specific allowed characters that have meaning in the URL string.

2.3 Keyword Removal (Signature Based Detection)

Input filtering describes the process of validating all incoming data. “Suspicious” input that might contain a code injection payload is either rejected, encoded, or the “offensive” parts are removed using so called “removal filters”. The protection approach implemented by these filters relies on removing predefined keywords.

Different kinds of attacks have different keywords. For example, SQL injection attack has the keywords like SELECT, INSERT, UPDATE, DELETE, UNION, etc.XSS attack has the keywords like<script, javascript, or document. Directory Traversal attack has the keywords like “dir, cmd, windows, ../”, etc.

Firstly, attack types are detected by their respective keywords in the system. This process is also known as keyword detection for each attack type.

3 Implementation of the System

3.1 Request Length Module

The length of the input requests to the web server can be used to detect the occurred attacks. If μ is considered to be the average length of n requests received by an application with the parameters of L1,L2,L3,……..,Ln in which Li represents the length of the received requests of i, and σ2 will be the variance of these requests, then according to the equation 1, the possibility of P for a request with the length of L will be as the following.

$$p = \frac{\sigma^2}{(l-\mu)^2} \tag{1}$$

The values of μ and σ2 are calculated separately in the education phase according to the received requests. After calculating these values in the test phase and while considering the pre-defined values and the size of the newly received request, the value of P will be calculated and if it’s higher than a threshold, that request will be considered as an anomaly request. In figure 2, you can see an example of the normal and abnormal requests which can be detected by this module. This method can easily detect attacks like Directory Traversal and Buffer

overflow. Because these attack inherently, have request sizes larger than the normal size. [1]

3.2 Regular Expression Patterns

The goal of a regular expression is to match a certain expression within a lump of text. A regular expression pattern is usually enclosed within slashes ('/').Regular expressions enable a powerful, flexible, and efficient text processing. This system can analyze how attack log file occurred by using DVWA web server. By inputting some attack patterns from input box and by POST method, we can analyze how certain types of attacks occurred in web server log file.

Compute Possibility

id	URL	mean	sigma	possibility	Result
1	/	65.93	5215.44	1.24	is normal
2	icons/poweredby.png	65.93	5215.44	2.37	is normal
3	icons/apache_pb.gif	65.93	5215.44	2.37	is normal
4	favicon.ico	65.93	5215.44	1.73	is normal
5	/	65.93	5215.44	1.24	is normal
6	adduser.php	65.93	5215.44	1.73	is normal
7	/adduser_submit.php	65.93	5215.44	2.37	is normal
8	/adduser.php	65.93	5215.44	1.79	is normal
9	/login_submit.php	65.93	5215.44	2.18	is normal
10	/login.php	65.93	5215.44	1.67	is normal
11	/login.php?op=login	65.93	5215.44	2.37	is normal
12	register.php	65.93	5215.44	1.79	is normal
13	register.php?op=reg	65.93	5215.44	2.37	is normal
14	/?_test1=c:\\windows\...	65.93	5215.44	.32	is normal
15	/register.php?regna...	65.93	5215.44	1.60	is normal
16	/register.php?regna...	65.93	5215.44	26.35	is normal
17	/register.php?regna...	65.93	5215.44	7.67	is normal
18	/?mode=a)/)/)/)/)/)/)/)...	65.93	5215.44	.26	is normal
19	/register.php?regna...	65.93	5215.44	1.09	is normal
20	/?mode=RihEE</->Ri...	65.93	5215.44	82.91	is attack
21	/register.php?regna...	65.93	5215.44	1218.39	is attack
22	/index.php?view=../../.....	65.93	5215.44	141.60	is attack
23	/?mode=<!--#exec c...	65.93	5215.44	36.64	is normal
24	/index.php?view=Oel...	65.93	5215.44	.12	is normal
25	/index.php?view=%n...	65.93	5215.44	.23	is normal
26	/index.php?view=../../.....	65.93	5215.44	315.01	is attack
27	/index.php?view=39" ...	65.93	5215.44	4.27	is normal
28	vulnerabilities/xss_s/ ...	65.93	5215.44	4.27	is normal
29	/index.php?view=%S...	65.93	5215.44	6.23	is normal

Fig. 2 Attack Detection with Request Length Module

In this system regex patterns for three web attacks and normal (attack free) patterns are predefined. Some patterns of regular expression used in this system are as follows:

[a-z A-Z]*/[a-z]+_s/ [A-Z]+/1.1=XSS
/((\%3C)|<)((\%2F)|\/)*[a-z0-9\%]+((\%3E)|>)/ix =XSS
/((\%3C)|<)((\%69)|i|(\%49))((\%6D)|m|(\%4D))((\%67)|g|(\%47))[^\n]
+((\%3E)|>)/I=XSS

/(javascript|vbscript|expression|applet|script|embed|object|iframe|frame|
frameset)/i=XSS
/(java)?script/i=XSS
/(\.|(%|%25)2E)(\.|(%|%25)2E)(\/|(%|%25)2F|\\|(%|%25)2C)/=Directory
Traversal
/exec(\s|\+)+(s|x)p\w+/ix=DirectoryTraversal
/(\||%00|system\(|eval\(|’|\\)/i=Directory Traversal
/(\’)|(\%27)|(\-\-)|(#)|(\%23)/ix=SQL Injection
/((\%3D)|(=))[^\n]*((\%27)|(\’)|(\-\-)|(\%3B)|(;))/i=SQL Injection
/\w*((\%27)|(\’))(\s|\+|\%20)*((\%6F)|o|(\%4F))((\%72)|r|(\%52))/ix
=SQLInjection
/((\%27)|(\’))(select|union|insert|update|delete|replace|truncate)/ix=SQL
Injection

By computing regex pattern analysis, the system can show the result as in figure 3.

Log File Analysis

Id	URL	Class
1	/	Normal
2	vulnerabilities/xss_s/ HTTP/1.1	XSS
3	icons/poweredby.png	Normal
4	icons/apache_pb.gif	Normal
5	favicon.ico	Normal
6	/	Normal
7	adduser.php	Normal
8	/adduser_submit.php	Normal
9	/adduser.php	Normal
10	/login_submit.php	Normal
11	/login.php	Normal
12	/login.php?op=login	Normal
13	register.php	Normal
14	register.php?op=reg	Normal
15	/?_test1=c:\\windows\\system32\\cmd.exe&_test2=/etc/passwd&_test3=|/bin/sh&_test4=(SELECT * FROM nonexistent) --&_test5=>/no/such/file&_test6=<scri...	SQL Injection
16	/register.php?regname=/../../../../../../../../../etc/passwd&regemail=vega@example.com&regpass1=vega&regpass2=vega	Normal
17	/register.php?regname=Joey&regemail=vega%nd%nd%nd%nd%nd%nd%nd%nd&regpass1=vega&regpass2=vega	Normal
18	/register.php?regname=Joey&regemail=vega@example.com&regpass1=vega&regpass2=..\\..\\..\\..\\..\\..\\..\\..\\..\\..\\..\\..\\etc\\passwd	Normal
19	/register.php?regname=Joey&regemail=vega@example.com&regpass1=vega&regpass2=vega	Normal
20	/register.php?regname=&regemail=&regpass1=&regpass2=14" AND "14"="15	Normal
21	/?mode=a))/	Normal
22	/?mode=<!--#exec cmd="echo -n vdpnA;echo -n RgYii" -->	SQL Injection
23	/?mode=RihEE</->RihEE/*RihEE"RihEERihEE'RihEERihEE`RihEE =	Normal
24	/index.php?view=%n%n%n%n%n%n%n%n%n%n%n%n%n%n%n%n%n%n%n%n%n%n%n%n%n%n%n%n%n%n%n%n%n%n%n%n%n%n%n%n%n%...	Normal
25	/index.php?view=../../../../../../../../../../../etc/passwd	Normal
26	/index.php?view=39" OR "39"="39	Normal
27	/index.php?view=OeIDgRjsoUmBpGpwarOJdIbGBYNsGaWakrlxqSvDaXlsvqGcayWjClqYBgWSQgLJVXBxxPNzlCqSDFWEnWcGMPEsXuDInsdrlMHesIJVbbVxb...	Normal
28	/index.php?view=%SYSTEMROOT%\win.ini	Normal
29	/index.php?view=../../../../../../../../../../../boot.ini	Normal

Fig. 3 Attack Classification with Regex Pattern Analysis

3.3 System Flow Diagram

Input to the system is web server log file. After preprocessing, web log data are stored in the database. URL decoding is performed on the data. First step of request length module is computed. The next step, regex pattern analysis is performed and attack detection results are produced.

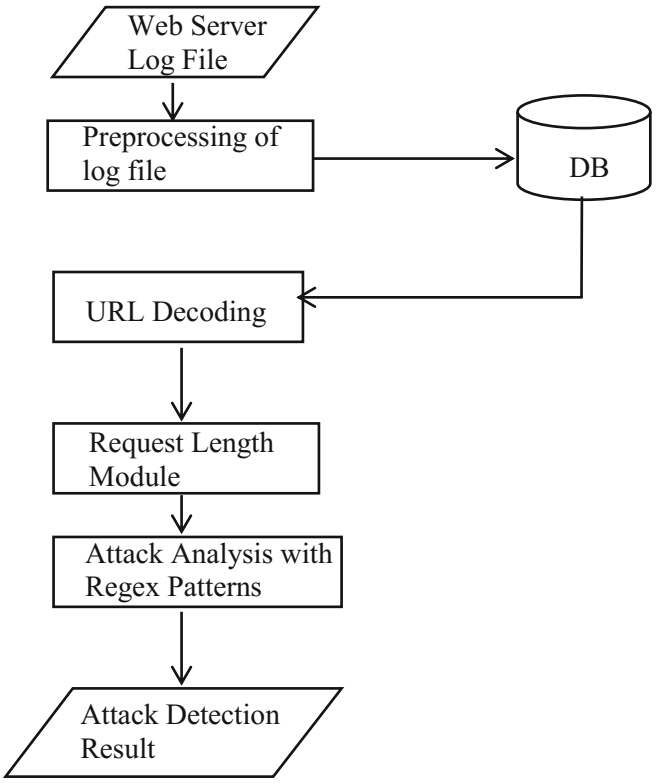


Fig. 4 System Flow Diagram

4 Conclusion and Experimental Results

This system presents about analyzing and classifying web application attacks. Combination of request length module and regular expression patterns are used in this system. Three types of attacks namely; SQL injection, XSS and directory traversal can be effectively classified by this system. Other attacks in the dataset that are not covered by this system will be resulted as unknown attacks. By computing request length module, unseen attacks can be detected. Predefined regex pattern analysis cannot be covered in some cases. For these condition, the system results as unknown attacks.

Detection rate of request length module and regex pattern analysis for each attack type is shown in figure 5. By analyzing this result, directory traversal attack can be more effectively detected than the other two attacks in request length module. By regex pattern analysis, SQL injection attack detection rate is higher than the others. Request length module is effective for unknown attacks. For known attacks and with certain patterns, regex pattern analysis is an effective method. Regex patterns in this system are for three types of attacks and normal access patterns. The experimental results are computed based on the dataset received by DVWA.

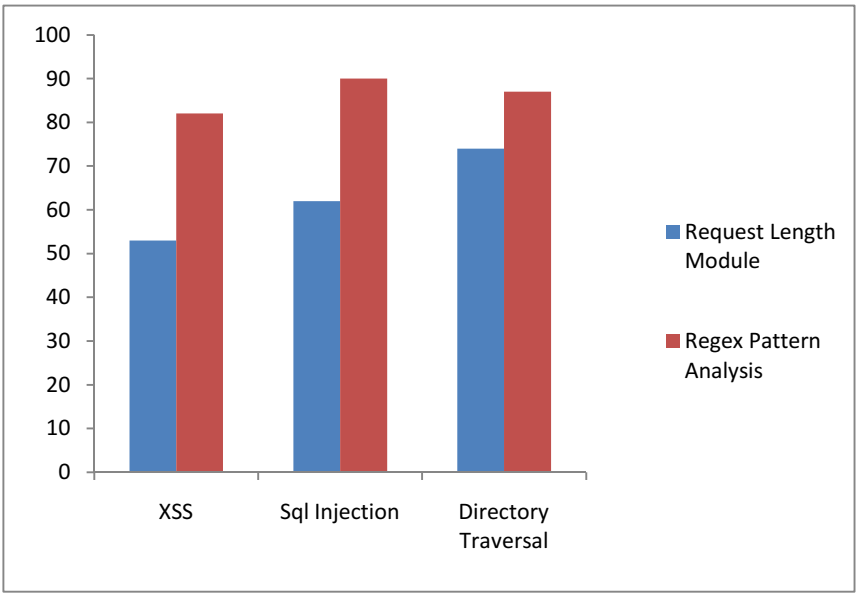


Fig. 5 Detection Rate of Three Attacks by Request Length Module and Regex Pattern Analysis

References

1. Vamsidhar, T., Ashok, R., Venkat, R.: Intrusion Detection System For Web Applications With Attack Classification. Journal of Global Research in Computer Science (2012)
2. Kruegel, C., Vigna, G., Robertson, W.: A multi-model approach to the detection of web-based attacks. Reliable Software Group. University of California, Santa Barbara (2005)
3. Meyer, R., Cid, C.: Detecting Attacks on Web Applications from Log Files. SANS Institute (2008)
4. Kruegel, C., Vigna, G.: Anomaly detection of Web-based attacks. In: Proceedings of the 10th ACM Conference on Computer and Communication Security(CCS 2003), Washington, DC, October 2003, pp. 251–261. ACM Press, New York (2003)

5. Mookhey, K.K., Burghate, N.: Detection of SQLInjection and CrosssiteScriptingAttacks (2004). http://www.blackhat.com/presentations/bhusa04/bhus04mookhey/old/bhus04mookhey_whitepaper.pdf
6. Robertson, W., Vigna, G., Kruegel, C., Kemmerer, R.: Using generalization and characterization techniques in the anomaly based detection of web attacks. In: 13th Annual Network and Distributed System Security Symposium, San Diego (2006)
7. Gallagher, B., Eliassi-Rad, T.: Classification of http attacks: A studyon the ecml/pkdd 2007 discovery challenge (2009)
8. Faradzhullaev, R.: Analysis of Web Server Log Files and Attack Detection. Institute of Information Technologies, Academy of Sciences of Azerbaijan (2007)

A Study on the Effects of Virtualization on Mobile Learning Applications in Private Cloud

Si Si Mar Win, Hnin Mya Aye and Than New Aung

Abstract Real time communication applications including Mobile learning application can be integrated with other software applications into one platform and deployed in private clouds to reduce capital expenditure and lower overall costs of daily based maintenance and real estate required for computer hardware. As a critical component of private clouds, virtualization may adversely affect a real time communication application running in virtual machines as the layer of virtualization on the physical server adds system overhead and contributes to capacity lose. Virtualization in the mobile can enable hardware to run with less memory and fewer chips, reducing costs and increasing energy efficiency as well. It also helps to address safety and security challenges, and reduces software development and porting costs. This study will investigate how to build an effective learning environment for both the University and learners by integrating the virtualization, private cloud technology and mobile learning applications.

Keywords Private cloud · Mobile learning · Virtualization

1 Introduction

Nowadays, mobile devices such as smart phone, tablet pcs, etc. are increasingly becoming an essential part of human life as the most effective and convenient communication tools not bounded by time and place. Mobile users accumulate rich experience of various services from mobile applications (e.g. iPhone apps, Google apps, etc.), which run on the devices and/or on remote servers via wireless networks. The rapid progress of mobile computing [10] becomes a powerful trend in the development of IT technology as well as commerce and industry fields. However, the mobile devices are facing many challenges in their resources

S.S.M. Win(✉) · H.M. Aye · T.N. Aung
University of Computer Studies, Mandalay, Myanmar
e-mail: {sisimarwin,hninmyaaye26,mdytna}@gmail.com

T.T. Zin et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 388,
DOI: 10.1007/978-3-319-23207-2_17

(e.g., battery life, storage, and bandwidth) and communications (e.g., mobility and security) [4]. The limited resources significantly impede the improvement of service qualities. The traditional mobile learning applications have limitations in educational resources. Cloud based mobile learning applications are introduced to overcome these limitations.

Cloud Computing technology has brought great opportunities to these technological learning environments. It has been driven by technological innovations as collaboration software, service oriented architectures (SOAs), and data center virtualization.

In cloud computing, virtualization can be a significant factor in performance loss of software applications because of interactions with the underlying virtual machine monitor and other virtual machines. Virtual computing resources, however, may influence the performance of software applications as adding a layer of virtualization software on the computer server increases overhead to the overall system and contributes to capacity loss. There has been an arising interest to examine the impact of virtualization on performance loss. The reason behind this interest is to do with the growing adoption of clouds computing by organizations which expect existing software applications to be able to run on virtual machines and to perform as good as on physical servers.

Cloud computing using virtualization technology has been widely recognized as the next generation's computing infrastructure. It offers some advantages by allowing users to use infrastructure (e.g., servers, networks, and storages), platforms (e.g., middleware services and operating systems), and software (e.g., application programs) provided by cloud providers (e.g., Google, Amazon, and Sales force) at low cost. In addition, Cloud computing enables users to elastically utilize resources in an on-demand fashion.

Mobile cloud computing brings new types of services and facilities for mobile users to take full advantages of cloud computing. As a result, mobile learning applications can be rapidly provisioned and released with the minimal management efforts or service provider's interactions.

2 Related Work

Mobile cloud computing is a model for transparent elastic augmentation of mobile device capabilities via ubiquitous wireless access to cloud storage and computing resources, with context-aware dynamic adjusting of offloading in respect to change in operating conditions, while preserving available sensing and interactivity capabilities of mobile devices.

The adoption of cloud computing using virtualization has unique security and privacy implications in mobile information systems. These aspects are related to ensuring that the data and processing controlled by a third party is secure and remains private, and the transmission of data between the cloud and the mobile device is secured [1]. Clouds provide access to data, but the challenge is how to ensure that only authorized entities can access the data. This requires a

combination of technical and non-technical means, i.e. clients need to trust their providers and the providers need to ensure their technical competence and integrity. This approach provides highest degree of control over performance, reliability and security. Using the cloud principles for virtual private cloud applications makes organizations better prepared to migrate or overflow to a public cloud provider when needed.

An approach of using virtual machine (VM) technologies executing the computation intensive software from mobile device is presented in [3]. In this architecture, a mobile user exploits VMs to rapidly instantiate customized service software on a nearby cloudlet and uses the service over WLAN. A cloudlet is a trusted, resource rich computer or a cluster of computers well connected to the Internet and available for use by nearby mobile devices. Rather relying on a distant cloud, the cloudlets eliminate the long latency introduced by wide-area networks for accessing the cloud resources. As a result, the responsiveness and interactivity on the device are increased by low-latency, one-hop, high bandwidth wireless access to the cloudlet. The mobile client acts as thin client, with all significant computation occurring in a nearby cloudlet. This approach relies on technique called dynamic VM synthesis. [10]

An E-Learning Mobile Device Enabled Ubiquitous E-learning Environment in cloud by utilizing a smart mobile application combined with the Quick Response Code mechanism was presented in [18]. It aimed to develop an interactive live mobile learning system, which streams live lectures to student's mobile devices with interacting facilities, using a mixture of existing and self-developed codec's. Alternatively, Stojan Kitanov presented a new model of mobile distance learning system (MDL) in an extended mobile cloud computing environment (MCC) by using High Performance Computing (HPC) Cluster Infrastructure, as well as some existing videoconferencing technologies enriched with mobile and wireless devices. Although described that this MCC model can be applied everywhere where there is need of fast and intensive computing and analysis of huge amount of data, such as modeling of 3D graphics visualization and animation in ecology, global climate solutions, financial risks, healthcare and medical learning, decoding genome projects, it cannot be performed by a conventional PC, Laptop or Mobile Device [17].

Similarly, a model for mobile learning system in virtualized cloud computing environment using high performance computing (HPC) and graphical processing unit (GPU) cluster infrastructure is presented in [15]. This model aggregated the power with new technologies to implement M-leaning process more effective with high performance and quick response.

3 Architecture of MCC

Cloud computing is an emerging area of distributed computing that offers many potential benefits to organizations by making information technology (IT) services available as a commodity. When they contract for cloud services, such as

applications, software, data storage, and processing capabilities, organizations can improve their efficiency and their ability to respond more quickly and reliably to their customers' needs. At the same time, there are risks to be considered, including maintaining the security and privacy of systems and information, and assuring the wise expenditure of IT resources.

Some of the leaders cloud computing leaders began in the early 2000s. Amazon Web Services (AWS) began in 2002, which provided a suite of cloud-based services including storage and computation. Amazon then launched its Elastic Compute cloud (EC2) [Amazon] in 2006. This combined with Amazon's Simple Storage Service (Amazon S3) [13] allowed customers to rent computer hardware so that they can run their own applications and store data in the cloud. Other cloud leaders include Microsoft, Google, and Apple. Eucalyptus [Eucalyptus, 2011b] provides private cloud solutions that are compatible with AWS, giving users an open source option for creating local clouds. Each of these vendors is developing cloud technologies with different goals in mind. Each has different company backgrounds and different target markets they wish to saturate.

Cloud computing is not a single type of system, but it encompasses a range of underlying technologies and configuration options. The strengths and weaknesses of the different cloud technologies, configurations, service models, and deployment methods should be considered by organizations evaluating services to meet their requirements. Cloud computing now offers organizations more choices regarding how to run infrastructures, save costs, and delegate liabilities to third-party providers. It has become an integral part of technology and business models, and has forced businesses to adapt to new technology strategies.

Accordingly, the demand for cloud computing has forced the development of new market offerings, representing various cloud service and delivery models. These models significantly expand the range of available options, and task organizations with dilemmas over which cloud computing model to employ.

3.1 Virtualization

Virtualization is the creation of a virtual (rather than actual) version of something, such as an operating system, a server, a storage device or network resources. There are three areas of IT where virtualization is making headroads, network virtualization, storage virtualization and server virtualization. Using virtualization technology, we could provide as little as 0.1 CPU in a virtual machine to the end user, therefore drastically increasing the utilization potential of a physical server to multiple users.

Additionally, virtualization is another key technology. It can maximize resource utilization efficiency and reduce cost of IaaS platform and user usage by promoting physical resource sharing. The dynamic migration function of virtualization technology can dramatically improve the service availability and this is attractive for many users.

A growing trend in the IT world is implementing virtualized servers. That is, software can be installed allowing multiple instances of virtual servers to be used. In this way, a dozen virtual servers can run on one physical server. Virtualization allows multiple operating systems to run on a single hardware device at the same time by more efficiently using system resources, like processors and memory. It enables pooling of computing resources from clusters of servers and dynamically assigning or reassigning virtual resources to applications on-demand. This study mainly focuses on the server virtualization on private cloud.

3.2 Virtual Infrastructures of Private Clouds (Virtual Private Clouds)

Virtual machines are a fundamental component of virtual infrastructures. While a virtual machine represents the virtual version of the hardware resources of an entire computer, a virtual infrastructure represents the interconnected hardware resources of an entire IT infrastructure including computers, database, network devices and shared storage resources. Organizations of all sizes build virtual server and desktop infrastructures to improve the availability, security and manageability of mission critical applications through a virtual infrastructure.

A virtual infrastructure shares physical resources of multiple machines across the entire infrastructure. A virtual machine shares the resources of a single physical computer across multiple virtual machines for maximum efficiency. Physical resources can be shared across multiple virtual machines and applications. The multiple servers of the underlying hardware infrastructure are aggregated along with networks and storage into a shared pool of IT resources that can be utilized by the applications when and where they're needed. This resource optimization drives greater flexibility in the organization and results in lower capital and operational costs. In a private cloud, an organization sets up a virtualized environment on its own servers, either in its own data centers or in those of a managed services provider. This structure is useful for organizations that either have significant existing IT investments or feel they absolutely must have total control over every aspect of their infrastructure (Reese, 2009).

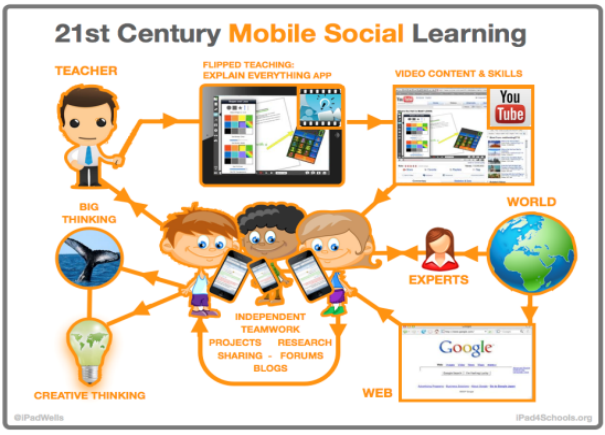
The key advantage of private clouds is control. In a private cloud, organizations retain full control over their infrastructure, but also gain some benefits of cloud computing such as the ability to reduce costs without the pitfalls, capitalizing on data security, corporate governance, and reliability concerns (Foley, 2008). Organizations are interested in private clouds because, in many instances, they cannot host their data outside of their firewalls due to privacy and legal issues, but they want to take advantage of the cloud computing architecture. Many of them want to remain in control of their systems and information and have already invested in hardware and software, the cost of which cannot be recovered (Linthicum, 2009). Private cloud architecture consists of two major components: a base hypervisor and a management server. A base hypervisor creates a layer of abstraction between virtual servers and the underlying hardware. The hypervisor are installed

directly on every physical server planned for creating a host and partition it into multiple virtual machines that can run simultaneously, sharing the physical resources of the underlying server. Each virtual machine represents a complete system, with processors, memory, networking, storage and BIOS, and can run an unmodified operating system and applications. The management server is responsible for centralized management of private cloud hosts and their virtual machines.

Virtual private clouds allow service providers to offer unique services to private cloud users. These services allow customers to consume infrastructure services as part of their private clouds. The ability to augment a private cloud, with on-demand and at-scale characteristics, is typical of a virtual private cloud infrastructure. Private cloud customers can seamlessly extend the trust boundaries (security, control, service-level management, and compliance) to include virtual private clouds.

3.3 Mobile Cloud Computing

The term mobile cloud computing was introduced not long after the concept of "cloud computing" launched in 2007. It has been attracting the attentions of entrepreneurs as a profitable business option that reduces the development and running cost of mobile applications, of mobile users as a new technology to achieve rich experience of a variety of mobile services at low cost, and of researchers as a promising solution for green IT [3].



Source:http://inthecloud.gjmueller.com/post/54446407492/how-mobile-social-learning-really-works

Fig. 1 Typical MCC Based Learning System

Mobile cloud computing (MCC) generally involves three components: mobile computing, cloud computing and wireless networks. MCC aims to enhance computational capabilities of resource-constrained mobile devices towards rich

and increasing user experience. MCC equips business and education sectors the opportunities for mobile network operators as well as cloud providers. More comprehensively, MCC can be defined as a rich mobile computing technology that influences united flexible resources of diverse clouds and networks technologies toward absolute asks , storage, and mobility to serve a multitude of different mobile devices anywhere, anytime over the special channel of Ethernet or Internet regardless of heterogeneous suitable environments and platforms based on the pay-as-you-use principle might including consumer, enterprise, trains coding , end-to-end security, home gateways, and mobile broadband-enabled services. Thus, MCC is defined as an expansion of cloud computing with a new ad-hoc infrastructure which depends on a mobile device [7].

4 Impact of Virtualization

The advances in virtualization, automation and distributed computing have allowed corporate data centers to become service providers that can meet the needs of customers outside their corporate boundaries. Using the cloud principles for private cloud applications makes organizations better prepared to migrate or overflow to a public cloud provider when needed.

The first step of evaluation is to consider the nature, the amount, and the transfer rate of the data and applications, how critical they are, and what are the minimum required levels of performance and availability. After that, another important characteristic to investigate is the data localization or distribution across the network.

Using virtualization, it is possible to achieve a redundancy system for all the services running on a data center. This new approach to high availability allows to share the running virtual machines over the servers up and running, by exploiting the features of the virtualization layer: start, stop and move virtual machines between physical hosts.

The evolution to cloud computing has advanced rapidly over the last few years. As a critical component of private clouds, however, virtualization may influence real time communication applications because adding a layer of virtualization software on the computer server adds overhead to the overall system. There has been an increasing interest to examine the impact of virtualization on performance loss. The reason behind this increased interest is to do with the growing adoption of clouds computing by organizations which expect existing software applications to be able to run on virtual machines and to perform as good as on physical servers.

A Spiceworks survey conducted in 2013 found that nearly two-thirds of SMBs had adopted server virtualization solutions, while 62 percent of SMBs had implemented cloud computing in some capacity. Jay Hallberg, said SMBs worldwide are making major investments in virtualization, cloud computing and mobile devices such as smartphones and tablets.

The use of cloud computing, with its dynamic scalability and virtualized resources usage, can empower mobile Learning by eliminating some of weaknesses of the mobile handheld devices, creating mobile Learning as a Service (mLaaS), focusing on the following four features: transparency; collaboration, extended into intra-organizational sharing of educational and learning resources; personnel learning; and motivational effects [8].

Since real time communication services require a certain level of system performance and availability to address communication latency and overhead bottleneck, it is essential to investigate potential performance implications of virtual private clouds on mobile learning applications.

Looking ahead, the impact of cloud computing and virtualization shows little signs of dissipating. A Host Review report by Brent Johnson asserted that these technologies are the future of the IT market. This report described that these tools allow organizations to rely less on employees and more on the Internet and automation. This point is especially attractive for Universities that may be struggling to control their overhead costs, but still want to experience the advantages of innovative technology and may be performed off-site and virtually. Universities that provide the live lectures and the proper training courses or presentations will remain successful over the long run, while those universities that do not will be left behind.

5 Conclusion

The findings of this study indicate that the real time communication application suffered from performance loss with a lower factor in private clouds than in a non-virtualized environment with comparative network capacity and computing resources. Server virtualization creates some significant challenges to real time communications in private clouds due to interactions between the underlying virtual machine monitor and other virtual machines. Adding a layer of virtualization software on the computer server adds overhead to the overall system. Therefore, virtual infrastructure is desired to reduce system capacity loss and application performance degradation and bring more cost savings to adoption of private clouds.

Moreover, given a common web service registry on the cloud, the performance benefits accrued by applications include hardware reuse, remote maintenance of hosted servers, storage provisioning and network provisioning with interoperability, scalability, and security. The application's transaction semantics can be reliably represented using some of the cloud infrastructure capabilities like server virtualization. Therefore mobile learning applications can be replicated easily on a cloud resulting in economies of scale. Using the enabling technologies enumerated, mobile learning applications can be deployed as web services to provide interoperability, business continuity, transaction persistence and server provisioning.

References

1. Rudenko, A., Reiher, P., Popek, G.J., Kuenning, G.H.: Saving portable computer battery power through remote process execution. Journal of ACM SIGMOBILE on Mobile Computing and Communications Review (1998)
2. Avanada: Global Survey of Cloud Computing, March 22, 2010
3. Jung, E., Wang, Y., Prilepov, I., Maker, F., Liu, X., Akella, V.: User-profile-driven collaborative bandwidth sharing on mobile phones. In: 1st ACM Workshop on Mobile Cloud Computing & Services: Social Networks and Beyond (MCS), No. 2 (2010)
4. Gao, H., Zhai, Y.: System design of cloud computing based on mobile learning. In: 3rd International Symposium on Knowledge Acquisition and Modeling (KAM), pp. 293–242, November 2010
5. IDA, New Delhi, India: Public vs Private vs Hybrid vs Community Cloud Computing: A Critical Review. I.J. Computer Network and Information Security, pp. 20–29, March 2014
6. Li, J.: Study on the development of mobile learning promoted by cloud computing. In: 2nd International Conference on Information Engineering and Computer Science (ICIECS), December 2010
7. Ali, M.: Green cloud on the horizon. In: Jaatun, M.G., Zhao, G., Rong, C. (eds.) Cloud Computing. LNCS, vol. 5931, pp. 451–459. Springer, Heidelberg (2009)
8. Alabbadi, M.M.: Mobile learning (mLearning) based on cloud computing: mLearning as a service (mLaaS). In: The Fifth International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies (2011)
9. Satyanarayanan, M.: Fundamental challenges in mobile computing. In: 5th annual ACM Symposium on Principles of Distributed Computing, pp. 1–7, May 1996
10. Satyanarayanan, M.: Mobile computing: the next decade. In: 1st ACM Workshop on Mobile Cloud Computing & Services: Social Networks and Beyond (MCS), June 2010
11. Chen, X., Liu, J., Han, J., Xu, H.: Primary exploration of mobile learning mode under a cloud computing environment. In: International Conference on E-Health Networking, Digital Ecosystems and Technologies (EDT), June 2010
12. Zhou, T.: Exploring mobile user acceptance based on UTAUT and contextual offering. In: International Symposium on Electronic Commerce and Security (2008)
13. http://aws.amazon.com/s3/
14. Shunye, W., Dayong, L., Zijuan, Z.: E-Learning system architecture based on Private Cloud for university. Journal of Chemical and Pharmaceutical Research **6**(5), 492–498 (2014)
15. Balghosoon, A.O., Rizwan Jameel Qureshi, M.: A Novel Model for Cloud Based Mobile Learning System. I.J. Information Engineering and Electronic Business **6**, 40–46 (2014)
16. Ghadirli1, H.M., Rastgarpour, M.: A Paradigm for the Application of Cloud Computing in Mobile Intelligent Tutoring System
17. Kitanov, S., Davcev, D.: Mobile cloud computing environment as a support for mobile learning. In: The Third International Conference on Cloud Computing, GRIDs, and Virtualization, Cloud Computing (2012)
18. Yeh, W.-T., Yang, C.-T.: Construct a Mobile Device Enabled Ubiquitous E-learning Environment in Cloud. ISBN: 978-0-9891305-4-7 ©2014 SDIWC

Developing Mobile Application Framework by Using RESTFul Web Service with JSON Parser

Ei Ei Thu and Than Nwe Aung

Abstract Nowadays, mobile devices offer new ways for users to access information. Web service can be built by using two separate ways: standard SOAP based and RESTFul web service. This paper presents the motivations and technical choices for creating RESTFul API integrated with mobile application. This application framework easy to deploy, test, maintain and rely on scalable and easily integrated infrastructure. And also explain why choose REST rather than SOAP and why choose JSON parser rather than XML.

Keywords Web Service · RESTFul · JSON

1 Introduction

Mobile web service provisioning is substantially expanding on the concept of 'anywhere, anytime and on any device' to a new paradigm ubiquitous mobile computing. It is used to improve access to meaningful, quickly and required information and content through mobile web services. Many of the problems of mobile web services can be solved by targeting the distributed nature and isolated deployment of mobile applications. One of the most promising ways to create viable web services for mobile devices is to add extra intelligence to the web services, both on the web service provider and the web service consumer. Mobile devices with their hardware limitations are generally not suitable to use Internet Services via Web Pages. The separations of user interface and service logic offered by Web Services are a new chance to bring internet services to mobile devices. Applications running on mobile devices, providing access to Web Services, can thereby be adapted to the specific device capabilities. To integrate Web Service technologies in mobile devices one has to consider the restrictions of these devices and the mobile communication system.

E.E. Thu(✉) · T.N. Aung
University of Computer Studies Mandalay (UCSM), Mandalay, Myanmar
e-mail: {eieithuet,mdytna}@gmail.com

T.T. Zin et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 388,
DOI: 10.1007/978-3-319-23207-2_18

Mobile Technology has now come up with "Libraries in Hand" trend. Our librarians are in move to determine how these devices are affecting information access and ensure that they are communicating with patrons and providing web content in the most appropriate and effective ways. Our Librarians must be prepared to take this challenge and put his efforts to increase the market and demand for mobile access to personalized facts and information anytime, anywhere on one's own handheld device. Web Services can be classified into two main categories: RESTFul and SOAP-based Web Services. This classification is based on the architectural style used in the implementation technology. SOAP stands for Simple Object Access Protocol. It is an object oriented technology that defines a standard protocol used for exchanging XML-based messages. REST stands for Representational State Transfer; it is a resource oriented technology that consists of a set of design criteria that define the proper way for using web standards such as HTTP and URIs. Although REST is originally defined in the context of the Web, it is becoming a common implementation technology for developing web services. RESTFul Web Services are implemented with Web standards (HTTP, XML and URI) and REST principles. REST principles include addressability, uniformity, connectivity and stateless. RESTFul Web Services are based on uniform interface used to define specific operations that are operated on URL resources [3].

The rest of the paper is organized as follow: In section 2, related works are introduced; this includes introductions to XML vs. JSON, web services in mobile devices and some android based applications. Section 3 briefly introduces multitire application of Web API, RESTFul Web Service and JSON parser. Section 4 presents overview system architecture and implemented screen shots. Finally, section 5 concludes this work.

2 Related Works

Varun Goyal [7], This paper described various aspects of web services in mobile devices, i.e. what are the limitations of mobile devices, connectivity issues, how to optimize the web service, comparing different protocols and frame work that can be used, performance analysis of SOAP and RESTFul web services, various libraries that can be used to create web services.

Anil Dudhe, etc. [4] analyzed the performance of SOAP and RESTFul web service in cloud environment. They have run and collected the results of REST and SOAP web service on Google App Engine 1.8.2. They showed that REST web services take less time for responding data by comparing the tested results.

Feda AlShahWan, etc. [5] showed that using a REST-based framework leads to a better performing offloading behavior, compared to SOAP-based mobile services. Distributed mobile services based on REST consume fewer resources and achieve better performance compared to SOAP based mobile services.

Dunlu Peng, etc. [6] investigated how to employ JSON as the data exchange format for web service applications. They compared with XML, using JSON-style data for exchanging can improve the performance of web service applications.

Their experimental results showed that JSON performs better than XML in being parsed, being serialized and being deserialized.

Isak Shabani, Besmir Sejdiu [8] implemented MyParking android application that helps users to find parking lots depending on their location. This application is executed in Android mobile platform and which accesses the SOAP Web services server.

Sarawut Markchit [9] proposed offering library resources system for web-based and mobile application with SOAP web services. Author developed web-based application with ASP.NET and mobile application with HTML5 and JQuery.

3 Web Service Technology

A web service is a method of communication between two or more electronic devices over the World Wide Web. W3C defines web services as a "software system designed to support interoperable machine-to-machine communication over a network". It has a network described in a machine process able format. Other systems can communicate with the web service in a manner recommended by its description using SOAP messages, typically transferred using HTTP with an XML or JSON serialization in conjunction with other Web-related standards [10]. Web services are platform neutral and generally text based which can developed, run and accessed on heterogeneous technologies. So they are interoperable.

3.1 Web API

Web API is a development in Web services where emphasis has been moving to simpler representational state transfer (REST) based communications. RESTFul APIs may not require XML based Web service protocols (SOAP and WSDL) to support their interfaces. RESTFul web APIs or RESTFul web service is a web API

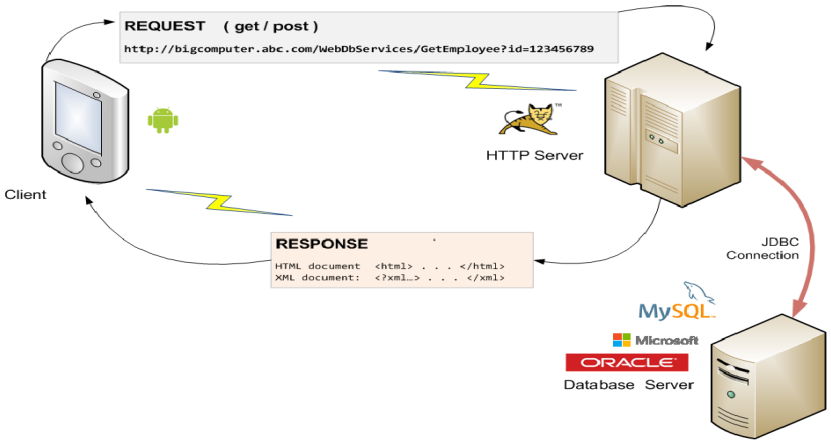


Fig. 1 Multi-tire application with application server and database server

implemented using HTTP and basis of REST. RESTFul API separates user interface involved from data storage. It improves flexibility of interface over multiple platforms and simplifies server components by making them stateless. Each request from client comprises all the state information and server does not hold client context in the session. Figure 1 illustrates the consuming web service for multi-tire application with application server and database server.

3.2 RESTFul Web Service

REST is a software application architecture modeled after the way data is represented, accessed, and modified on the web. It is an architectural style for distributed hypermedia systems. In the REST architecture, data and functionality are considered resources, and these resources are accessed using Uniform Resource Identifiers (URIs), typically links on the web. The resources are acted upon by using a set of simple, well defined operations. The REST architecture is fundamentally client-server architecture, and is designed to use a stateless communication protocol, typically HTTP. In the REST architecture, clients and servers exchange representations of resources using a standardized interface and protocol. These principles encourage REST applications to be simple, lightweight, and have high performance. RESTFul web services are web applications built upon the REST architecture. They expose resources (data and functionality) through web URIs, and use the four main HTTP methods to create, retrieve, update, and delete resources. RESTFul web services typically map the four main HTTP methods to the so-called CRUD actions: create, retrieve, update, and delete [1]. Figure 2 shows the RESTFul web services architecture.

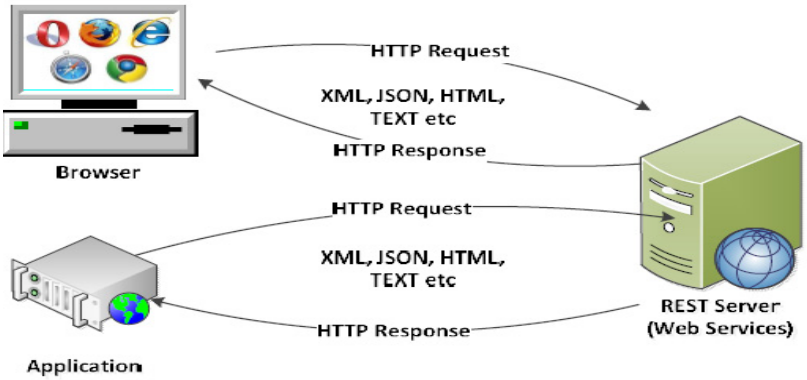


Fig. 2 RESTFul Web Service Architecture

3.3 JSON (JavaScript Object Notation) Parser

For the past few years, XML web services have dominated the arena for web services, as XML was touted as the ubiquitous medium for data exchange. However,

using XML as the medium for data payload suffers from the following problems: XML representation is inherently heavy. The use of opening and closing tags add a lot of unnecessary weight to the payload. XML representation is difficult to parse. While on the desktop, the DOM (Document Object Model) and SAX (Simple APIs for XML) are the two commonly used method for parsing XML Documents; on the mobile platform using DOM and SAX are very expensive, both computationally and in terms of memory requirements.

In recent years, another data interchange format has been gaining in popularity - JSON, or JavaScript Object Notation. JSON is a lightweight, text-based, language-independent data interchange format. It was derived from the ECMAScript (European Computer Manufacturers Association) programming language, but is programming language independent. JSON defines a small set of structuring rules for the portable representation of structured data. Like XML, JSON is a text-based open standard for representing data, and it uses characters such as brackets "[{]}", colon ":" and comma ",", to represent data. Data are represented using simple key/value pairs, and more complex data are represented as associative arrays. JSON is agnostic about numbers. In any programming language, there can be a variety of number types of various capacities and complements, fixed or floating, binary or decimal. That can make interchange between different programming languages difficult. JSON instead offers only the representation of numbers that humans use: a sequence of digits. All programming languages know how to make sense of digit sequences even if they disagree on internal representations. That is enough to allow interchange [2]. The following figure 3 shows the applying JSON parser in proposed work.

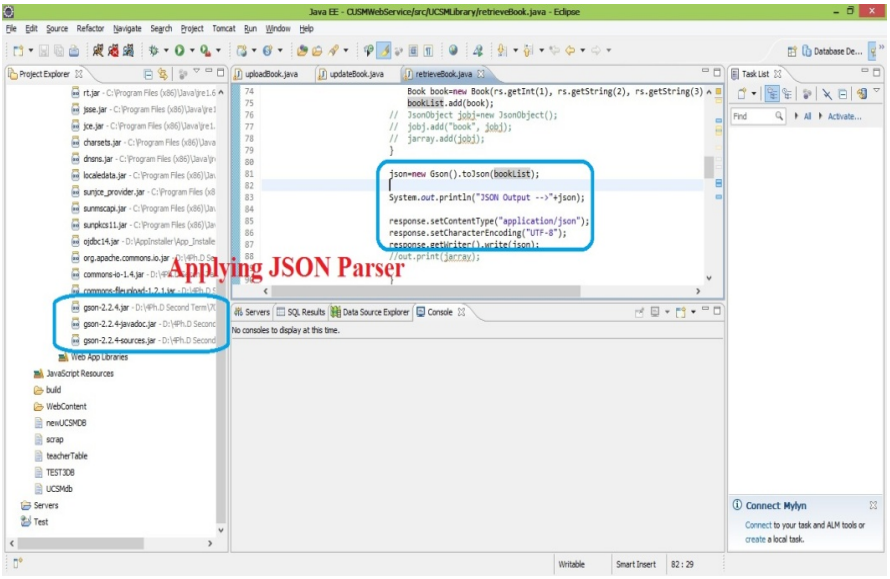


Fig. 3 Applying JSON Parser in Proposed Application

4 System Analysis

System consists of two mobile application modules: Library management module and University student and staff information management module. Student and staff information module can perform CRUD (create, read, update, delete) action for staff and student information. This application intends to use for university student affair and manage department. Library management module offers to support for librarian, student, teacher and staff. In this module, librarian also performs CRUD action for e-book and can also create unique user ID for user. Firstly, the user needs to sign up to use the library application. This library module will automatically check the signing up user is teacher or student or staff by accessing information from student and staff manage module. And then the system will automatically create unique user ID according to their occupation (teacher, student, and staff). Because the librarian needs to classify access permission for each user. User will be access e-book according to their permission. So this application framework can support even librarian in offline. And also provide interoperability and transparently exchanging information through RESTFul web service by using proposed two application module. These two applications can easily integrate to university's existing wireless network by changing http protocol. So that this proposed work can provide efficient and usable mobile network infrastructure for university environment. Figure 4 shows the proposed mobile network infrastructure.

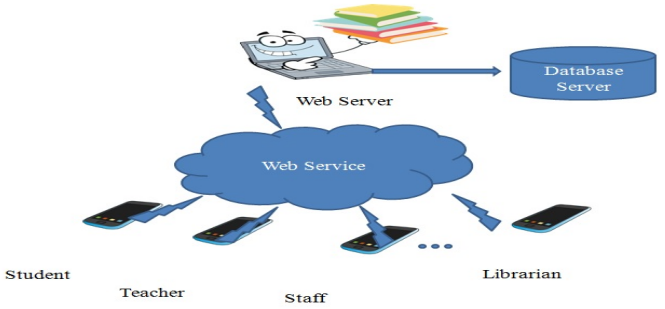


Fig. 4 Mobile Application Network Infrastructure

4.1 Testing

The proposed mobile framework developed with java based android programming language for mobile app and server side implemented with RESTFul technology based java servlet programming language. The proposed work implement RESTFul web service and deployed on Apace Tomcat 7.0 web server. And two mobile applications implemented using Android Developer Tool (ADT) bundle, Android 4.2.2-API level 18 and runs on Android Emulator. To parse the multimedia and text format data through web service using gson-2.2.4 and apache-mime4j-core.The following figure 5 shows the testing two mobile apps on android emulator.

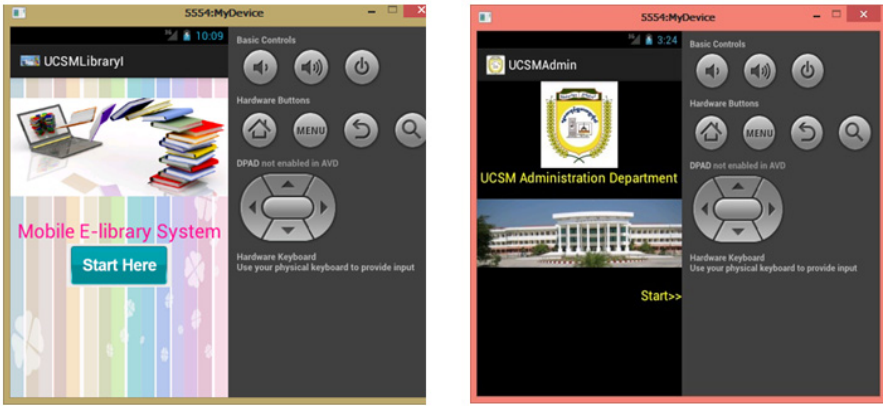


Fig. 5 Testing on Android Emulator

5 Conclusion

The processing capabilities of mobile devices have increased enormously in the recent years. This paper aims to develop the RESTFul web service to access e-book from university library with mobile network framework. The proposed system implemented android based mobile library infrastructure and tested successfully using RESTFul web service provisioning concept. The proposed work can support efficient mobile library framework with usability and interoperateability. At the present, the proposed mobile network framework includes two application modules; in the future this framework can easily integrate with other application module. And also, the proposed work can extend as a mobile learning framework within university and can also implement with other web service technology and other parser.

References

1. Bohara, M.H., Mishra, M.: RESTFul Web Service Integration using Android Platform, pp. 1–6. DAIICT, Gandhinagar (2013)
2. ECMA-262 (ISO/IEC 16262), ECMAScript® Language Specification, 3rd edn., October 2013
3. Hamad, H., Saad, M., Abed, R.: Performance Evaluation of RESTful Web Services for Mobile Devices. International Arab Journal of e-Technology **1**(3), January 2010. Computer Engineering Department, Islamic University of Gaza, Palestine
4. Dudhe, A., Sherekar, S.S.: Performance Analysis of SOAP and RESTFul Mobile Web Service in Cloud Environment. International Journal of Computer Applications (2014). Department of Advanced Software and Computing Technologies, Pune, India
5. Alshahwan, F., Moessner, K.: Evaluation of Distributed SOAP and RESTFul Mobile Web Services. International Journal on Advances in Networks and Services (2010). Centre for Communications Systems Research, University of Surrey, UK

6. Peng, D., Cao, L., Xu, W.: Using JSON for Data Exchanging in Web Service Applications. Journal of Computational Information Systems (2011). School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, China
7. Goyal, V.: Web Services in Mobile Devices. Computer Science Department, Rochester Institute of Technology (2013)
8. Shabani, I., Sejdiu, B.: Consuming Web Services on Android Mobile Platform for Finding Parking Lots. IJACSA **6**(2) (2015). University of Prishtina, Republic of Kosovo
9. Markchit, S.: Offering Library Resources through Web-site and Mobile Systems with Web Services for Central Library Suratthani Rajabhat University. World Journal of Computer Application and Technology **3**(1) (2015). Faculty of Science and Technology, Suratthani Rajabhat University, Thailand
10. Web Service Glossary. W3C. (retrieved, April 2015)

Part III
High Speed Computation and Applications in Information Systems

Subquadratic Space-Complexity Parallel Systolic Multiplier Based on Karatsuba Algorithm and Block Recombination

Chiou-Yng Lee, Che Wun Chiou and Jim-Min Lin

Abstract Recently, high-performance elliptic curve cryptography has gained great attention for resource-constrained applications. In this paper, we use (a, b)-way Karatsuba algorithm to derive a new way of k-way Karatsuba algorithm and block recombination (KABR) approach. We have derived a new parallel systolic multiplication with subquadratic space complexity based on k-way KABR approach. By theoretical analysis, it is shown that the proposed structure using k-way BRKA has significantly less computation delay, less area-delay product, and less area. Moreover, the proposed structure can provide the desired tradeoff between space and time complexity.

1 Introduction

Elliptic curve cryptosystem (ECC) is the most popular public-key protocols, due to its key-length is shorter than the well-known RSA with the same level security. The ECC has an advantage feature of applications to be suitable for resource-constrained applications, such as smart cards, telephones, and cell phones. We realize efficient ECC applications [5, 16], which depend on point multiplication on elliptic curves, which involves several point additions on elliptic curves. The implementation of point additions can be realized by either projective coordinates or affine coordinates over binary field $GF(2^m)$ or prime field $GF(p)$. Projective coordinates are suitable for high-performance ECC designs, since each point addition involves additions,

C.-Y. Lee(✉)
Lunghwa University of Science and Technology, Guishan, Taiwan
e-mail: pp010@mail.lhu.edu.tw

C.W. Chiou
Chien Hsin University of Science and Technology, Taoyuan, Taiwan

J.-M. Lin
Feng Chia University, Taichung 40724, Taiwan

T.T. Zin et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 388,
DOI: 10.1007/978-3-319-23207-2_19

squarings, and multiplications but does not involve inversion operation. In the binary extension fields, addition and squaring are fast operations and involve significantly less area, while multiplication is the bottleneck of cryptographic algorithms due to its large area and time complexity. Therefore, the efficient hardware design for point multiplication in resource-constrained environments requires high-performance multiplication, which involves highly computation time, less area, and less area-delay product.

$GF(2^m)$ multiplication is widely studied on hardware architectures and software implementations. Its hardware architecture depends on the field generated by irreducible polynomials. In ANSI X9.62 [2] and NIST (FIPS 186-2) [3] standards, trinomials and pentanomials in binary extension fields are recommended to be used for the implementation of elliptic curve digital signature algorithm (ECDSA). Up to date, the hardware implementation of finite field multiplication is classified into systolic and non-systolic architectures. Systolic architecture has an advantage feature of high-performance ECC applications to be suitable for resource-constrained applications, since it can provide high-throughput and fast computation. Many of efficient parallel multiplier structures have been proposed in [9, 10, 21].

The two-way divide and conquer algorithm, referred to as Karatsuba algorithm (KA), was introduced in [8] and has $O(m^{1.59})$ space complexity, while the naive multiplication involves $O(m^2)$ space complexity. The generalized k-way divide and conquer scheme is suggested by Toom and Cook [4, 18], in particular, when $k = 3$, its space complexity has $O(m^{1.47})$. During recent few years, KA algorithm has received more attention to implement bit-parallel multipliers with subquadratic space complexity bit-parallel multipliers [6, 17, 19]. In order to solve the problems of successive multiplication in some applications, e.g., multiplicative inversion, exponentiation, and point multiplication, three-operand multiplications according to recursive KA and TMVP decompositions are suggested [11, 13]. The generalized (a, b)-way KA approaches with $a \neq b$ are presented in [12, 14, 15] for exploring subquadratic space complexity digit-serial multipliers.

In this paper, we use the (a, b)-way KA decomposition and block recombination to explore a novel k-way KABR approach. We have derived a new parallel systolic multiplication with subquadratic space-complexity systolic architecture according to k-way KABR approach, while traditional systolic multipliers are based on grade-school computation approaches. By theoretical analysis results, the proposed k-way KABR multiplier has less area, less computation time, and less area-time product compared to the existing systolic architectures. Moreover, the proposed k-way KABR multiplier can provide the desired tradeoff between space and time complexities for parallel multiplication architecture.

2 Review of (a, b)-Way Karatsuba Algorithm

A univariate polynomial $A = a_0 + a_1 x + \cdots + a_{n-1} x^{n-1}$ over GF(2) can be transformed into a bivariate polynomial as

$$A(x, y) = \sum_{j=0}^{p-1} \sum_{i=0}^{q-1} a_{i,j} x^i y^j \tag{1}$$

where

$$a_{i,j} = a_{i+qj}$$

$$n = pq$$

$$y = x^q$$

We can use $A = A(x, x^q)$ to perform the basis conversion from bivariate polynomial to univariate polynomial, which is a free of hardware cost. Based on bivariate and univariate polynomials, (a, b)-way KA decomposition with different orders is introduced in [12]. We briefly review a (4, 2)-way and (6,3)-way KA decompositions as follows.

2.1 (4,2)-Way KA

Let us consider two polynomials $A = A_0 + A_1 x + A_2 y + A_3 xy$ and $B = B_0 + B_1 x$. Let us denote that the symbol "A_{ij}" is "$A_i + A_j$". The product of A and B can be obtained as

$$\begin{aligned} AB &= A_0B_0 + (A_{01}B_{01} + A_0B_0 + A_1B_1)x + A_2B_0y \\ &+ (A_{23}B_{01} + A_2B_0 + A_3B_1)xy + A_1B_1x^2 + A_3B_1x^2y \\ &= C_0 + C_1x + C_2y + C_3xy + C_4x^2 + C_5x^2y. \end{aligned} \tag{2}$$

The product AB in (2) uses 6 partial products to compute C_0, C_1, C_2, C_3, C_4, and C_5, while the naive multiplication requires 8 partial products to compute the product AB. Based on (2), Fig. 1 shows the high-level architecture for the (4,2)-way KA decomposition. It involves two evaluation point (EP1 and EP2) units, one point-wise product (PWM) unit, and reconstruction (R) unit. Four units according to (2) are defined as

$$\begin{cases} P_A = EP1(A) = (A_0, A_1, A_2, A_3, A_{01}, A_{23}) \\ P_B = EP2(B) = (B_0, B_1, B_{01}) \\ W = PWM(P_A, P_B) = (A_0B_0, A_1B_1, A_2B_0, A_3B_1, A_{01}B_{01}, A_{23}B_{01}) \\ \quad = (W_0, W_1, W_2, W_3, W_4, W_5) \\ C = R(W) = (W_0, W_{014}, W_2, W_{235}, W_1, W_3) \end{cases} \tag{3}$$

The decomposition of (3) can be performed recursively to implement polynomial multiplication. Each multiplication is transformed into 6 partial products of the digits of A and B whose degrees are reduced to, respectively, about quarter and half.

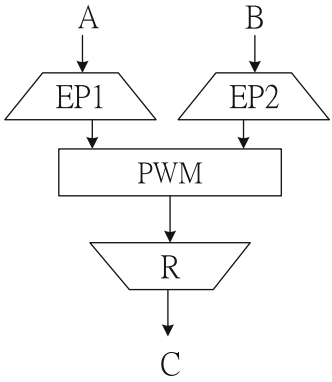


Fig. 1 High-level architecture for (4,2)-way KA decomposition

If the decomposition algorithm is terminated after degeneration of polynomials into single-bit coefficients, polynomial A is required to be of $n = 4^i$-bits for $i > 1$, and polynomial B is required to be of $n^{\log_4 2} (= 2^i = \sqrt{n})$ bits. Let "S" and "D" denote the number of gates and delays, respectively. The algorithm in (2) involves 6 partial products and $(1.5n + 2.5n^{\log_4 2} - 4)$ additions. The critical path of (2) is given by $D(\frac{n}{4}) + 3T_X$. Therefore, the complexity of the (4,2)-way KA decomposition is given by the following recurrence relations.

$$\begin{cases} S_{\otimes}(n) \leq 6S_A(\frac{n}{4}), S_A(1) = 1 \\ S_{\oplus}(n) \leq 6S_{\oplus}(\frac{n}{4}) + 1.5n + 2.5n^{\log_4 2} - 4, S_X(1) = 0 \\ D(n) \leq D(\frac{n}{4}) + 3T_X, D(1) = T_A \end{cases} \tag{4}$$

We can obtain that the complexity bound given by (4) is estimated to be

$$\begin{cases} S_{\otimes}(n) \leq n^{\log_4 6} \\ S_{\oplus}(n) \leq \frac{69}{20} n^{\log_4 6} - 3n - \frac{5}{4} n^{\log_4 2} + \frac{4}{5} \\ D(n) \leq T_A + 3(\log_4 n) T_X \end{cases} \tag{5}$$

2.2 (6,3)-Way KA

Let A and B be two polynomials of the forms $A = A_0 + A_1 x + A_2 x^2 + A_3 y + A_4 xy + A_5 x^2 y$ and $B = B_0 + B_1 x + B_2 x^2$, respectively. Polynomial A is 6-term bivariate polynomial, and polynomial B is 3-term univariate polynomial. The product $C = AB = C_0 + C_1 x + C_2 x^2 + C_3 y + C_4 xy + C_5 x^2 y + C_6 x^3 + C_7 x^4 + C_8 x^3 y + C_9 x^4 y$ can obtain the following formula

$$\begin{cases} C_0 = W_0, C_1 = W_0 + W_1 + W_6 \\ C_2 = W_0 + W_1 + W_2 + W_7 \\ C_3 = W_3, C_4 = W_3 + W_4 + W_8 \\ C_5 = W_3 + W_4 + W_5 + W_9 \\ C_6 = W_1 + W_2 + W_{10}, C_7 = W_2 \\ C_8 = W_4 + W_5 + W_{11}, C_9 = W_5 \end{cases} \tag{6}$$

where

$$\begin{cases} W_0 = A_0B_0, W_1 = A_1B_1, W_2 = A_2B_2, W_3 = A_3B_0, \\ W_4 = A_4B_1, W_5 = A_5B_2, W_6 = A_{01}B_{01}, W_7 = A_{02}B_{02}, \\ W_7 = A_{02}B_{02}, W_8 = A_{34}B_{01}, W_9 = A_{35}B_{02}, \\ W_{10} = A_{12}B_{12}, W_{11} = A_{45}B_{12}. \end{cases}$$

The product C can be derived from 12 partial products W_is according to (6). Applying this strategy, in each iteration multiplication is transformed into 12 partial products, where the degrees of A and B are reduced by a factor of about one sixth and one third, respectively. If the decomposition algorithm is terminated after degeneration of polynomials into single-bit coefficients, then the degrees of two polynomials A and B are required to be $n = 6^i$ and $n^{\log_6 3}$, respectively. According to the recursive formula in (6), the algorithm involves 12 sub-products and $(\frac{10}{3}n + \frac{17}{3}n^{\log_6 3} - 14)$ additions. We have the following recursive expressions of computational complexity.

$$\begin{cases} S_{\otimes}(n) \le 12S_{\otimes}(\frac{n}{6}),\ S_A(1) = 1 \\ S_{\oplus}(n) \le 12S_{\oplus}(\frac{n}{6}) + \frac{10}{3}n + \frac{17}{3}n^{\log_6 3} - 14, S_X(1) = 0 \\ D(n) \le D(\frac{n}{6}) + 3T_X,\ D(1) = T_A \end{cases} \tag{7}$$

To solve the above recursive relations, we can find the complexity bound of (6,3)-way KA as

$$\begin{cases} S_{\otimes}(n) = n^{\log_6 12} \\ S_{\oplus}(n) \le 3.95n^{\log_6 12} - \frac{10}{3}n - \frac{17}{9}n^{\log_6 3} + \frac{14}{11} \\ D(n) \le T_A + 3(\log_6 n)T_X \end{cases} \tag{8}$$

3 Proposed Systolic Multiplication over $GF(2^m)$

In the finite field $GF(2^m)$, a field element can be represented by the polynomial basis representation as $A = a_0 + a_1x + ... + a_{m-1}x^{m-1}$ over $GF(2)$. From Section 2, any polynomial A can be based on the bivariate polynomial representation in (1). Polynomial A is used by $y = x^2$ to transform two parts as $A = A_0 + A_1x$, where $A_0 = \sum_{i=0}^{q-1} a_{0,i}y^i$ and $A_1 = \sum_{i=0}^{q-1} a_{1,i}y^i$, where $a_{0,i} = a_{2i}$ and $a_{1,i} = a_{2i+1}$. Observing this representation, we can find that the coefficients of A_0 and A_1 are even-term and odd-term coefficients of A, respectively. This transformation is called the 2-way split method. In general, we can be extended by using $k-$way split method to transform the polynomial A as

$$A = \sum_{i=0}^{k-1} A_i x^i \tag{9}$$

where

$$A_i = \sum_{j=0}^{q-1} a_{ik+j} y^j,\ y = x^k,\ q = \left\lceil \frac{m}{k} \right\rceil.$$

Using k-way split method in (9), the proposed parallel multiplication algorithm and architecture is discussed as follows.

3.1 2-Way Split Method

Assume that the field $GF(2^m)$ is constructed from an irreducible polynomial $F(x)$. Let A, B and C be three elements in $GF(2^m)$, where $C = AB \bmod F(x)$. Using 2-way plit method with $y = x^2$, two polynomials A and B are represented as $A = A_0 + A_1 x$ and $B = B_0 + B_1 x$, respectively, where $A_i = \sum_{j=0}^{q-1} a_{i,j} y^j$ and $B_i = \sum_{j=0}^{q-1} b_{i,j} y^j$ for i=0 and 1. Thus, the product C can be re-expressed as

$$C = (A_0 + A_1 x)(B_0 + B_1 x) \bmod F(x) \tag{10}$$

$$= A_0 B_0 + (A_0 B_1 + A_1 B_0)x + A_1 B_1 x^2 \bmod F(x)$$

Next, let us consider that two polynomials B_0 and B_1 are grouped by d-digits, such as $B_i = \sum_{j=0}^{p-1} B_{i,j} y^{dj}$, where $B_{i,j} = \sum_{l=0}^{d-1} b_{i,dj+l} y^l$ and $p = \left\lceil \frac{m}{2d} \right\rceil$. The product C in (10) can be rewritten as

$$C = C_0 + C_1 y^d + \cdots + C_{p-1} y^{d(p-1)} \tag{11}$$

$$= ((C_{p-1}) y^d + C_{p-2}) y^d + \cdots) y^d + C_0 \bmod F(x)$$

where

$$C_i = A_0 B_{0,i} + (A_0 B_{1,i} + A_1 B_{0,i})x + A_1 B_{1,i} x^2$$

$$= A_0 B_{0,i} + A_1 B_{1,i} x^2 + (A_0 B_{1,i} + A_1 B_{0,i})x$$

Based on (a, b)-way KA decomposition in Section 2, its architecture in Fig. 1 involves four modules (EP1, EP2, PWM, and R modules). We can obtain the following property with block recombination approach:

Corollary 1. (block recombination) Assume that four sub-words as A_1, A_2, B_1, B_2, and $C = C_1 + C_2$, where $C_1 = A_1 B_1$ and $C_2 = A_2 B_2$. Based on the structure of the KA scheme, the product C can be recombined as follows:

Algorithm 1. Proposed multiplication scheme based on 2-way KABR

Inputs: $A = A_0 + xA_1$ and $B = B_0 + xB_1$ are two element in GF(2^m).
Output: $C = AB \; mod \; F(x)$.
1. $C = 0$, $D_{p-1} = 0$.
2. $B_i = \sum_{j=0}^{p-1} B_{i,j} y^{dj}$, where $B_{i,j} = \sum_{l=0}^{d-1} b_{i,dj+l} y^l$ for i=0 and 1, $y = x^2$, and $p = \lceil \frac{m}{2d} \rceil$.
3. $P_{A_0} = EP(A_0)$, $P_{A_1} = EP(A_1)$, $P_{B_{0,p-1}} = EP(B_{0,p-1})$, and $P_{B_{1,p-1}} = EP(B_{1,p-1})$
4. for $i = p - 1$ to 0
5. $C = Cy^d + D_i \bmod F(x)$.
6. $D_{i-1} = R(W_{0,i}) + x^2 R(W_{1,i}) + xR(W_{2,i} + W_{3,i})$, where $W_{0,i} = P_{A_0} \odot P_{B_{0,i}}$, $W_{1,i} = P_{A_1} \odot P_{B_{0,i}}$,
$W_{2,i} = P_{A_0} \odot P_{B_{1,i}}$, $W_{3,i} = P_{A_1} \odot P_{B_{0,i}}$
7. $P_{B_{0,i-1}} = EP(B_{0,i-1})$, $P_{B_{1,i-1}} = EP(B_{1,i-1})$
8. endfor
9. $C = Cy^d + D_0 mod \; F(x)$.

$$C = \mathrm{R}(W_1) + \mathrm{R}(W_2) = R(W_1 + W_2) \tag{12}$$

where

$$W_1 = P_{A_1} \odot P_{B_1},$$

$$W_2 = P_{A_2} \odot P_{B_2}.$$

Using (a,b)-way KA decomposition and block recombination (KABR) approach, the sub-product C_i in (11) can be expressed as

$$C_i = R(W_{0,i}) + x^2 R(W_{1,i}) + xR(W_{2,i} + W_{3,i}) \tag{13}$$

where

$$W_{0,i} = P_{A_0} \odot P_{B_{0,i}}$$

$$W_{1,i} = P_{A_1} \odot P_{B_{1,i}}$$

$$W_{2,i} = P_{A_1} \odot P_{B_{0,i}}$$

$$W_{3,i} = P_{A_0} \odot P_{B_{1,i}}$$

From the structure of KA decomposition, its architecture involves EP, PWM, and R components, and all component are independent computed. For this reason, assume that the partial product C_i in (13) is segmented into two-step computation as

Step-1: computes P_{A_0}, P_{A_1}, $P_{B_{0,i}}$, and $P_{B_{1,i}}$.
Step-2: computes $C_i = R(W_{0,i}) + x^2 R(W_{1,i}) + xR(W_{2,i} + W_{3,i})$.

Using two-step computation, Algorithm 1 shows the proposed multiplication using KABR approach. Fig.2 shows the proposed architecture for a new parallel systolic

multiplier based on Algorithm 1. It consists of three main parts, i.e., one pre-computation evaluation point (PCEP) cell, p processing element (PE) cells, and one final reduction-accumulator (FRAC) cell. In PCEP cell (as shown in Fig. 3c), A_0, A_1, $B_{0,p-1}$, and $B_{1,p-1}$ are, respectively, to go through 2 EP1 components and 2 EP2 components to generate P_{A_0}, P_{A_0}, $P_{B_{0,p-1}}$, and $P_{B_{0,p-1}}$ based on Step 3 of Algorithm 1. Each PE cell (as shown in Fig. 3a) consists of a PWM-R component, 2 EP2 components, and a RAC component. The PWM-R component is based on Fig. 3b to compute $D_{i-1} = R(W_{0,i}) + yR(W_{1,i}) + xR(W_{2,i} + W_{3,i})$ in Step 6, which consists of 4 PWM components, three component additions (CA1, CA2, CA3), two EP2 components, and three R components. The RAC component (as shown in Fig. 3d) is used to perform Step 5 for computing $C = Cy^d + D_i \bmod F(x)$, which consists of one accumulation (AC) module and one reduction polynomial (RP) module. Two EP2 components in Fig. 3a are used to perform Step 7 for computing $P_{B_{0,i-1}} = EP(B_{0,i-1})$, $P_{B_{1,i-1}} = EP(B_{1,i-1}))$. Note that those components in each PE are computed in parallel. The FRAC cell is used to perform the computation of Step 9, it structure is the same of the RAC component.

We follow the proposed structure in Fig. 2 for computing the multiplication $C = AB \bmod F(x)$. At first clock cycle, P_{A_0}, P_{A_0}, $P_{B_{0,p-1}}$, and $P_{B_{0,p-1}}$ are generated by PCVP cell, and stores in four registers ($< A0 >$, $< A1 >$, $< B0 >$, and $< B2 >$). During each clock cycle, two values $A0$ and $A1$ are still stored in two registers $< A0 >$ and $< A1 >$, and go through each PE cell to provide the computation of partial product D_i according to (13). After a latency of $(p + 1)$ cycles, we need extra one cycle to compute final reduction polynomial in the FRAC cell. Therefore, the proposed architecture requires $(p + 2)$ clock cycles (duration of each cycle is $T_A + (D^R(d) + 1)T_X$, where d is the selected digit-size, $D^R(d)$ is the delay of R component, and $p = \lceil \frac{m}{2d} \rceil$.

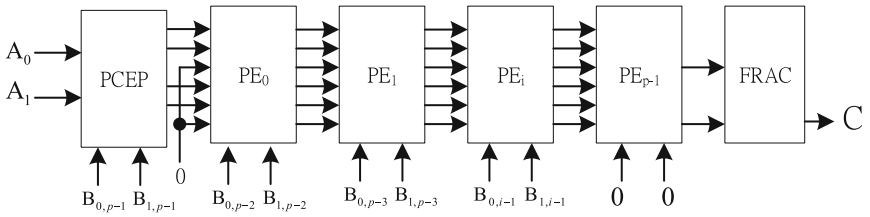


Fig. 2 The proposed parallel systolic multiplication architecture

3.2 *k*-Way Split Structure

Generally, we use k-way split method to split two polynomials as $A = \sum_{i=0}^{k-1} A_i x^i$ and $B = \sum_{i=0}^{k-1} B_i x^i$, where A_i and B_i are $\frac{m}{k}$-bit polynomials. Each of B_i is grouped by d-digits, such as $B_i = \sum_{j=0}^{p-1} B_{i,j} y^{dj}$, where $B_{i,j} = \sum_{l=0}^{d-1} b_{i,dj+l} x y^l$ and $p = \lceil \frac{m}{kd} \rceil$. Thus, the product C can be rewritten as

$$C = ((C_{p-1})y^d + C_{p-2})y^d + \cdots)y^d + C_0 \bmod F(x)$$

where

$$C_i = C_{i,0} + C_{i,1}x + \cdots + C_{i,2k-2}x^{2k-2}$$

$$C_{i,j} = \sum_{h+k=i} A_h B_{k,j}$$

Employing KABR approach, the partial product C_i can be given by

$$C_i = \sum_{j=0}^{2k-2} R\Big(\sum_{h+k=i} P_{A_h} \odot P_{B_{k,j}}\Big)x^j \tag{14}$$

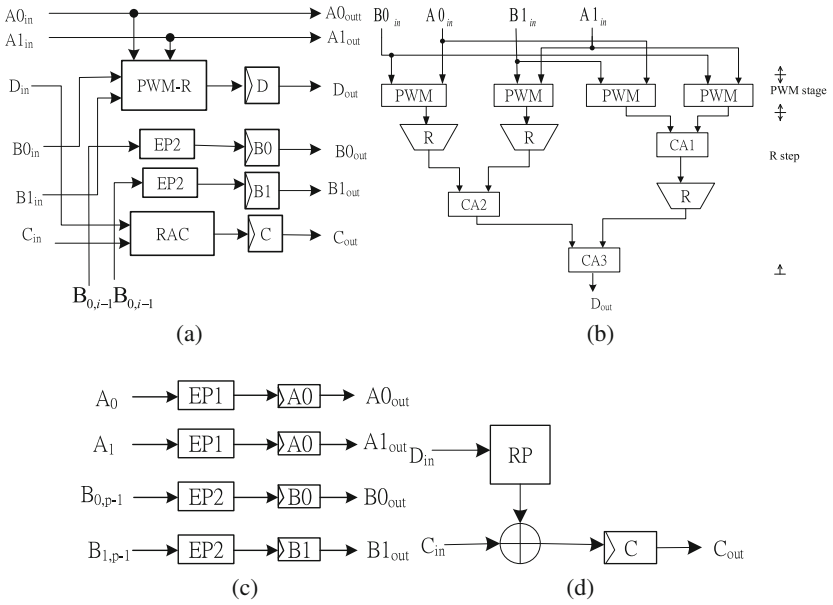


Fig. 3 (a) PE unit; (b) PWM-R unit; (c) PCEP unit; (d) RAC unit

Table 1 The complexities of three components in (4,2)-way and (6,3)-way KAs

Components	Complexities for (4,2)-KA	Complexities for (6,3)-way KA
EP	$S_{\oplus}^{EP1}(m) = \frac{6}{5}m^{log_4 6} - m$	$S_{\oplus}^{EP1}(m) = m^{\log_6 12} - m$
	$S_{\oplus}^{EP2}(m) = \frac{1}{4}m^{log_4 6} - \frac{1}{4}m^{log_4 2}$	$S_{\oplus}^{EP2}(m) = \frac{1}{3}m^{\log_6 12} - \frac{1}{3}m^{\log_6 3}$
	$D^{EP1}(m) = (\log_4 m)T_X$	$D^{EP1}(m) = (\log_6 m)T_X$
	$D^{EP2}(m) = (\log_4 m)T_X$	$D^{EP2}(m) = (\log_6 m)T_X$
PWM	$S_{\otimes}^{PWM}(m) = m^{\log_4 6}$	$S_{\otimes}^{PWM}(m) = m^{\log_6 12}$
	$D^{PWM}(m) = T_A$	$D^{PWM}(m) = T_A$
R	$S_{\oplus}^{R}(m) = \frac{11}{5}m^{log_4 6} - 2m - m^{log_4 2} + \frac{4}{5}$	$S_{\oplus}^{R}(m) = 2.62m^{log_6 12} - \frac{7}{3}m - \frac{14}{9}m^{log_6 3} + \frac{14}{11}$
	$D^{R}(m) = (2\log_4 m)T_X$	$D^{R}(m) = (2\log_6 m)T_X$

Based on Fig. 2, we can use (14) to realize the structure of PE cell. The proposed k-way KABR-based multiplier involves $\lceil \frac{m}{kd} \rceil + 2$ cycles, and duration of each cycle is $T_A + (1 + D^R(d) + \log_2 k)T_X$.

4 Area and Time Complexities

4.1 Complexities of Our Proposed Multiplier

From Section 2, Table 1 lists the complexity of each component for (4,2)-way and (6,3)-way KA decompositions. In the following, we use the complexities of each component in Table 1 to estimate the complexity of our proposed KABR-based systolic multiplier.

4.1.1 2-Way Split Structure

The parallel systolic structure of Fig. 2 for 2-way split method require one PCVP cell, p PE cells, and one FRAC cell, where $p = \lceil \frac{m}{2d} \rceil$. As shown in Fig. 3b, we can find that the complexity of CA1 perform the computation of $T_0 = W_{2,i} + W_{3,i}$, which involves $S_{\oplus}^{CA1}(\frac{m}{2}) = S_{\otimes}^{PWM}(\frac{m}{2})$ XOR gates and T_X delay. The complexity of CA2 thus performs $T_1 = R(W_{0,i}) + yR(W_{1,i})$ involves $S_{\oplus}^{CA2}(\frac{m}{2}) = (\frac{m}{2} + d - 2)$ XOR gates and T_X delay, since $R(W_{0,i})$ and $R(W_{1,i})$ are produced by $(\frac{m}{2} + d - 1)$-bit polynomials; and CA3 performs $T_1 + xR(T_0)$, which is a free of hardware cost. Assume that $S^B(d)$ is the number of output bits for EP2 component. For clarity, based on the case of (4,2)-way KA decomposition, we have obtained $S^B(d) = d^{\log_2 3}$. Similarly, for (6,3)-way KA decomposition, we have obtained $S^B(d) = d^{\log_3 6}$. PWM-R component in PE cell is based on Fig.3b to produce $(m + 2d - 2)$-bit polynomial, the register $< D >$ requires $(m + 2d - 2)$ 1-bit registers. We use the following Lemma 1 to estimate the complexity of RAC unit.

Lemma 1. *In Algorithm 1, the RAC is performed by $C = Cy^d + D_i \bmod F(x)$, where $y = x^2$. If $F(x)$ is an irreducible trinomial of the form $x^m + x^n + 1$, then we can find that the RAC involves $S_{\oplus}^{RAC}(m) = (m + 2d)$ XOR gates and m-bit register, and requires $3T_X$ delay.*

As stated above, we can find that the PCVP cell (as shown in Fig.3c) involves $(2S_{\oplus}^{EP1}(\frac{m}{2}) + 2S_{\oplus}^{EP2}(\frac{m}{2}))$ XOR gates and $2S_{\otimes}^{PWM}(\frac{m}{2}) + 2S^B(d)$ 1-bit registers, and requires $D^{EP1}(\frac{m}{2})$ T_X delay. Each PE cell (as shown in Fig.3a) involves $4S_{\otimes}^{PWM}(\frac{m}{2})$ AND gates, $(3S_{\oplus}^{R}(\frac{m}{2}) + 2S_{\oplus}^{EP2}(\frac{m}{2}) + S_{\oplus}^{RAC}(m) + S_{\oplus}^{CA2}(\frac{m}{2}) + S_{\oplus}^{CA1}(\frac{m}{2}))$ XOR gates, and $(2S^B(d) + 2m + 2d - 2)$ 1-bit registers, and requires $D^R(\frac{m}{2}) + T_A + T_X$ delay. The FRAC cell is equivalent to the complexity of RAC component. Therefore, the proposed structure using 2-partition scheme has the following time and space complexities:

$$\begin{cases} \#XOR = & 2S_{\oplus}^{EP1}(\frac{m}{2}) + 2(p+1)S_{\oplus}^{EP2}(\frac{m}{2}) + 3pS_{\oplus}^{R}(\frac{m}{2}) + pS_{\oplus}^{CA1}(\frac{m}{2}) + pS_{\oplus}^{CA2}(\frac{m}{2}) + (p+1)S_{\oplus}^{RAC}(m) \\ \#AND = & 4pS_{\otimes}^{PWM}(\frac{m}{2}) \\ \#FF = & 2S_{\otimes}^{PWM}(\frac{m}{2}) + m + 2(p+1)S^{B}(d) + p(2m+2d-2) \\ Latency = & p+2 \\ Delay = & D^{R}(\frac{m}{2}) + T_A + T_X \end{cases} \tag{15}$$

4.1.2 *K*-Way Structure

Using k-way split scheme, we find that, in each PE cell, PWM-R unit according to (14) involve $(k^2 - 2k + 1)$ CA1 components, k^2 PWM components, $(k-1)$ CA2 components, and $(2k-3)$ R components. Each of CA1 consists of $S_{\oplus}^{CA1}(\frac{m}{k}) = S_{\otimes}^{PWM}(\frac{m}{k})$ XOR gates and T_X delay. Each of CA2 involves $S_{\oplus}^{CA2}(\frac{m}{k}) = (\frac{m}{k} + d - 2)$ XOR gates and T_X delay. We can use Table 1 to estimate the complexity of EP1, EP2, PWM, and R components. Therefore, the proposed k-way KABR-based multiplier has the following time and space complexities:$\log_2 k$

$$\begin{cases} \#XOR = & kS_{\oplus}^{EP1}(\frac{m}{k}) + k(p+1)S_{\oplus}^{EP2}(\frac{m}{k}) + (2k-1)pS_{\oplus}^{R}(\frac{m}{k}) \\ & +(k^2 - 2k + 1)pS_{\oplus}^{CA1}(\frac{m}{k}) + (k-1)pS_{\oplus}^{CA2}(\frac{m}{k}) + (p+1)S_{\oplus}^{RAC}(m) \\ \#AND = & k^2 pS_{\otimes}^{PWM}(\frac{m}{k}) \\ \#FF = & kS_{\otimes}^{PWM}(\frac{m}{k}) + m + k(p+1)S^{B}(d) + p(2m + kd - 2) \\ Latency = & p+2 \\ Delay = & T_A + (\log_2 k + D_2^{R}(\frac{m}{k}))T_X \end{cases} \tag{16}$$

For clarity, using (4,2)-way KA decomposition, the proposed 4-way KABR-based multiplier can be found to have the following complexities

$$\begin{cases} \#XOR = & \frac{13}{6}m^{1+\log_4 3} + \frac{53}{15}m^{\log_4 6} - 3m^{0.5} + \frac{19}{32}m^{1.5} - \frac{5}{2}m - \frac{2}{5} \\ \#AND = & \frac{4}{3}m^{1+\log_4 3} \\ \#FF = & \frac{13}{6}m^{\log_4 6} + \frac{4}{3}m^{\log_4 3} + 2m + m^{1.5} - m^{0.5} \\ Latency = & m^{0.5} + 2 \\ Delay = & T_A + (1 + \log_4 m)T_X \end{cases} \tag{17}$$

4.2 Comparison of Area and Time Complexities

The area and time complexity in terms of logic gate count, register count, critical path delay, and latency of the proposed structure and the existing structures of [9, 20] is listed in Table 2. As shown in this table, our proposed structure in Fig.2 has $O(m^{1+\log_4 3})$ space complexity according 4-partition (4,2)-way KABR approach, while the existing parallel systolic multipliers have $O(m^2)$ space complexity. From this theoretical analysis, the proposed multiplier has subquadratic space complexity. Moreover, the proposed multiplier can lead low-latency complexity if the number of k-way split method is increased.

Table 2 Comparison of area-time complexity of parallel systolic multipliers

design	#AND	#XOR	#register	Latency	critical-path
[9]	m^2	$1.5m^2 + 0.5(m + n^2 - n)$	$4m^2 + m$	$m + 1$	$T_A + T_X$
[20]	m^2	$m^2 + 2m - 3$	$m^2 + 2m^{1.5} + 2m - 3m^{0.5}$	$m^{0.5} + 1 + log_4 m$	$2T_X$
[7]	$2m^2$	$2m^2$	$3m^2$	$m + 1$	$T_A + T_X$
Fig.2	$\frac{4}{3}n^{1+log_4 3}$	S_1	S_2	$2 + \frac{1}{2}n^{0.5}$	$T_A + (1 + \log_4 n)T_X$

note:(1) $S_1 = \frac{13}{6}n^{1+\log_4 3} + \frac{53}{15}n^{\log_4 6} - 3n^{0.5} + \frac{19}{32}n^{1.5} - \frac{5}{2}n - \frac{2}{5}$ and $S_2 = \frac{13}{6}n^{\log_4 6} + \frac{4}{3}n^{\log_4 3} + 2n + n^{1.5} - n^{0.5}$, where $n = 4^i$ for $i > 1$.
(2) Lee's multiplier in [9] is proposed based on trinomials $x^m + x^n + 1$.
(3) The proposed 4-way BRKA multiplier is based on (4,2)-way KA decomposition.

Table 3 Comparison of various subquadratic digit-serial multipliers over $GF(2^{223})$ in terms of latency (cycles), critical-path delay $T_{CPD}(ns)$, total critical delay $T_{TCD}(ns)$, area (μm^2), area-delay product (ADP)$(\mu m^2)ns$

Multipliers	Latency	T_{CPD}	T_{TCD}	Area	ADP
[9]	224	0.14	31.4	1,527,676	47,907,912
[20]	20	0.16	3.2	389,799	1,247,359
[7]	224	0.14	31.4	939,181	29,452,744
Fig.2	10	0.22	2.2	145,164	319,362

Note: The proposed 4-way BRKA multiplier is based on (4,2)-way KA decomposition.

We have used the NanGate's Library Creator and the 45-nm FreePDK Base Kit from North Carolina State University (NCSU) [1] to synthesize our proposed multiplier and the corresponding existing multipliers. Table 3 lists the comparison of area and time complexity, which is based on trinomial $F(x) = x^{223} + x^{33} + 1$. As shown in this table, we can find that the multiplier [20] is better than other multipliers. We use 4-way KABR approach to implement the parallel systolic multiplier. The proposed structure has has significantly less area and less area-delay product (ADP) compared to the corresponding existing multipliers. We also find that our proposed multiplier can obtain low-latency systolic architecture. Therefore, we can show that parallel systolic multiplier based on k-way KABR approach can achieve subquadratic space complexity. Our parallel systolic structure is suitable for high-performance ECC cryptographic processors for resource-constrained applications, while the existing multipliers involve larger area and larger ADP.

5 Conclusions

In this paper, we use the existing (a, b)-way KA decomposition and block recombination to derive a new way of k-way KABR approach. Based on this approach, we can achieve a subquadratic space-complexity parallel systolic multiplier, while the corresponding multipliers have $O(m^2)$ space complexity due to its designs are based on shoolbook computation approach. Moreover, based on the proposed structure,

we can obtain significantly less computation delay compared to the existing parallel systolic multipliers. From theoretical analysis, the proposed structure can be suitable for high-performance elliptic curve point multiplications in resource-constrained environments.

References

1. Nangate standard cell library. http://www.si2.org/openeda.si2.org/projects/nangatelib/
2. Public key cryptography for the financial services industry: The elliptic curve digital signature algorithm (ecdsa) (1999)
3. Digital signature standard (dss). Federal Information Processing Standards, Publication 186–2 (2000)
4. Cook, S.: On the minimum computation time of functions, master's thesis, Harvard University (1966)
5. El Gamal, T.: A Public Key Cryptosystem and a Signature Scheme Based on Discrete Logarithms. In: Blakely, G.R., Chaum, D. (eds.) CRYPTO 1984. LNCS, vol. 196, pp. 10–18. Springer, Heidelberg (1985)
6. Hasan, M.A., Meloni, N., Namin, A.H., Nègre, C.: Block recombination approach for subquadratic space complexity binary field multiplication based on toeplitz matrix-vector product. IEEE Trans. Computers **61**, 151–163 (2012)
7. Jain, S.K., Song, L., Parhi, K.K.: Efficient semisystolic architectures for finite-field arithmetic. IEEE Trans. VLSI Syst. **6**, 101–113 (1998)
8. Karatsuba, A.A., Ofman, Y.: Multiplication of multidigit numbers on automata. Soviet Physics Doklady **7**, 595–596 (1963)
9. Lee, C.-Y., Horng, J.-S., Jou, I.-C., Lu, E.-H.: Low-Complexity Bit-Parallel Systolic Montgomery Multipliers for Special Classes of GF(2^m). IEEE Trans. Computers **54**, 1061–1070 (2005)
10. Lee, C.-Y., Meher, P.K.: Efficient bit-parallel multipliers over finite fields GF(2^m). Computers & Electrical Engineering **36**, 955–968 (2010)
11. Lee, C.-Y., Meher, P.K.: Efficient subquadratic space complexity architectures for parallel MPB single- and double-multiplications for all trinomials using toeplitz matrix-vector product decomposition. IEEE Trans. on Circuits and Systems, 62-I (2015)
12. Lee, C.-Y., Meher, P.K.: Subquadratic space-complexity digit-serial multipliers over $GF(2^m)$ using generalized (a, b)-way karatsuba algorithm. IEEE Trans. on Circuits and Systems 62-I (2015)
13. C.-Y. Lee, P. K. Meher, and C.-P. Chang, Efficient m-ary exponentiation over $GF(2^m)$ using subquadratic ka-based three-operand montgomery multiplier. IEEE Trans. on Circuits and Systems 61-I, 3125–3134 (2014)
14. Lee, C.-Y., Meher, P. K., Lee, W.-Y.: Subquadratic space complexity digit-serial multiplier over binary extension fields using toom-cook algorithm. In: The International Symposium on Integrated Circuits (ISIC) (2014)
15. Lee, C.-Y., Yang, C.-S., Meher, B.K., Meher, P.K., Pan, J.-S.: Low-complexity digit-serial and scalable SPB/GPB multipliers over large binary extension fields using (b, 2)-way karatsuba decomposition. IEEE Trans. on Circuits and Systems 61-I, 3115–3124 (2014)
16. Lidl, R., Niederreiter, H.: Introduction to Finite Fields and Their Applications, 2nd edn. Cambridge University Press (1997)

17. Montgomery, P.L.: Five, six, and seven-term karatsuba-like formulae. IEEE Trans. Computers **54**, 362–369 (2005)
18. Toom, A.L.: The complexity of a scheme of functional elements realizing the multiplication of integers. Soviet Mathematics Doklady **3**, 714–716 (1963)
19. van der Hoeven, J., Lecerf, G.: On the complexity of multivariate blockwise polynomial multiplication. In: ISSAC 2012, pp. 211–218 (2012)
20. Xie, J., Meher, P.K., Mao, Z.: High-throughput finite field multipliers using redundant basis for FPGA and ASIC implementations. IEEE Trans. on Circuits and Systems 62-I, 110–119 (2015)
21. Xie, X.-N., Chen, G., Li, Y.: Novel bit-parallel multiplier for $GF(2^m)$ defined by all-one polynomial using generalized karatsuba algorithm. Inf. Process. Lett. **114**, 140–146 (2014)

Problems on Gaussian Normal Basis Multiplication for Elliptic Curve Cryptosystem

C.W. Chiou, Y.-S. Sun, C.-M. Lee, Y.-L. Chiu, J.-M. Lin and C.-Y. Lee

Abstract Several standards such as IEEE Standard 1363-2000 and FIPS 186-2 employ Gaussian normal basis (GNB). Gaussian normal basis is a special class of normal basis. Gaussian normal basis can solve the problem that multiplication in normal basis is an very difficult and complicated operation. Two equations have been proposed in the literature to transfer GNB to polynomial basis for easy multiplication. However, we find that GNB is not correctly transformed to polynomial basis for some m values over $GF(2^m)$. We will show the problems and expect some feedback about this problem from other researchers.

1 Introduction

Elliptic curve cryptosystem (ECC) [1, 2] is a powerful public-key cryptosystem for insuring information security of M-commerce on resource constrained smart phones. The arithmetic operations in $GF(2^m)$ have been largely applied in Elliptic curve cryptosystem and pairing-based cryptography [3]. ECC requires a smaller key size than RSA cryptosystem [4]. For example, ECC with 160-bit key has same security level as RSA with 1024-bit key. Therefore, ECC is suitable for resource constrained devices other than RSA. NIST and ANSI suggested finite fields for use in the ECDSA [5, 6]. The finite field multiplication is the most important arithmetic operation in $GF(2^m)$.

C.W. Chiou(✉) · Y.-S. Sun · C.-M. Lee · Y.-L. Chiu
Chien Hsin University of Science and Technology, Taoyuan City 32097, Taiwan
e-mail: {cwchiou,sunys,cmlee,B10013136}@uch.edu.tw

J.-M. Lin
Feng Chia University, Taichung City 407, Taiwan
e-mail: jimmy@fcu.edu.tw

C.-Y. Lee
Lunghwa University of Science and Technology, Taoyuan City 33306, Taiwan
e-mail: PP010@mail.lhu.edu.tw

T.T. Zin et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 388,
DOI: 10.1007/978-3-319-23207-2_20

Because, other complicated arithmetic operations such as exponentiation, division, and inversion can be computed by repeated multiplications. Therefore, it is important to explore efficient multiplier over large finite fields for resource-constrained devices. Efficiency of Multiplications in $GF(2^m)$ heavily depends on field element representations. There are three popular bases to represent field elements: polynomial basis (PB)[7]-[15], dual basis (DB) [16]-[21], and normal basis (NB) [22]-[35]. The major advantage of normal basis is its almost hardware-free squaring operation which can be easily carried out by cyclically shifting its binary representation. Thus, NB multipliers are very efficient in carrying out square operations in squaring, multiplicative inversion, and exponentiation operations. However, multiplication in normal basis is hardily realized. To overcome this problem, some special classes of normal basis have been presented to simplify normal basis multiplication. Optimal normal basis (ONB) [27] is one special class with the low-est space complexity in normal basis. But, only two types of ONB, type-1 and type-2, have been founded in the literature. Gaussian normal basis (GNB) is a special class of normal basis with low hardware complexity. All positive integers, except for those are divisible by eight, have GNB [36]. Both type-1 and type-2 ONB are same as type-1 and type-2 GNB, respectively. GNB now has been widely applied in several standards such as IEEE Standard 1363-2000 [5], FIPS 186-2 [37], ISO 11770-3 [38], and ANSI X9.62 [6]. Ash et al. [36] said that all positive integers except those are divisible by eight have GNB. As aforementioned, multiplication using GNB is hardly realized. Thus, GNB with type-t (t is an integer number) over $GF(2^m)$ is transformed to a polynomial basis with mt elements. In other words, PB transformed from GNB has t multiples of m elements in type-t GNB. However, some integer m values can not find sufficient t multiples. Two equations in the literature for computing t multiples of elements in GNB are applied for giving elements in PB from GNB. Results show that they both can not find sufficient t multiples of elements in GNB for some integer m values. This study will show this problem.

2 Background

There is always a normal basis $\psi = \{\beta^{2^0}, \beta^{2^1}, \cdots, \beta^{2^{m-1}}\}$ for a finite field $GF(2^m)$ for any positive integer m, where β is a normal element. Let any elements A and B in $GF(2^m)$ can be represented as $A = (a_0, a_1, \cdots, a_{m-1}) = \sum_{i=0}^{m-1} a_i \beta^{2^i}$ and $B = (b_0, b_1, \cdots, b_{m-1}) = \sum_{i=0}^{m-1} b_i \beta^{2^i}$ where a_i and $b_i \in GF(2)$ for $0 \leq i \leq m-1$. The major features of the normal basis are as follows:

Proposition 1. *Let A and B be two normal elements in $GF(2^m)$, we have obtained as*

1. $A^{2^r} = \sum_{i=0}^{m-1} a_{<i+r>_m} \beta^{2^i}$ for $0 \leq i \leq m-1$.
2. $A^{2m} = A$.
3. $(A+B)^2 = A^2 + B^2$.

Proposition-1 shows that the squaring of an element A in normal basis is just a right cyclic shift of its coordinates and it is almost hardware-free. The normal basis is termed the Gaussian normal basis with type-t (t is an integer and) if $p = mt + 1$ is a prime number and $gcd(mt/k, m) = 1$, where k is the multiplication order of 2 modulo p. It is noted that GNBs exist for any positive integer m, except that m is not divisible by eight. The GNB with type-t has the following properties:

$$\beta = \sum_{i=0}^{t-1} \gamma^{2^{mi}} \tag{1}$$

$$\gamma^{mt+1} = \gamma^{(mt+1) mod (mt+1)} = 1 \tag{2}$$

where γ is primitive $(mt + 1)$th root of unity in $GF(2^m)$. Then, β is called Gaussian period of type (m, t).

3 The Proposed Problems for GNB with Type-t

A Gaussian normal basis with type-t $\psi = \{\beta^{2^0}, \beta^{2^1}, \cdots, \beta^{2^{m-1}}\}$ can be transformed to a polynomial basis $\psi^* = \{\gamma^1, \gamma^2, \cdots, \gamma^{mt}\}$ using one of the following two equations:

$$\beta = \sum_{i=0}^{t-1} \gamma^{\tau^i}, \text{and } \tau^t = 1 \mod (mt + 1)[34] \tag{3}$$

$$\beta = \gamma + \gamma^{2^m} + \cdots + \gamma^{2^{(t-1)m}}[33] \tag{4}$$

Let any one element $A = (a_0, a_1, \cdots, a_{m-1}) = \sum_{i=0}^{m-1} a_i \beta^{2^i}$ belong to GNB ψ can be represented to $A = (a_0^*, a_1^*, \cdots, a_{mt}^*)$ in PB ψ^* using (3) as follows:

$$A^* = \sum_{i=0}^{m-1} a_i (\sum_{j=0}^{t-1} \gamma^{\tau^j})^{2^i} = \sum_{i=0}^{m-1} a_i \sum_{j=0}^{t-1} \gamma^{\tau^j 2^i} \tag{5}$$

$$= \sum_{i=0}^{m-1} (a_i \gamma^{\tau^0 2^i} + a_i \gamma^{\tau^1 2^i} + \cdots + a_i \gamma^{\tau^{t-1} 2^i})$$

The element A also can be represented as A^* using (4) as follows:

$$A^* = \sum_{i=0}^{m-1} (a_i \gamma + a_i \gamma^{2^m} + \cdots + a_i \gamma^{2^{(t-1)m}})^{2^i} \tag{6}$$

$$= \sum_{i=0}^{m-1} (a_i \gamma^{2^i} + a_i \gamma^{2^m 2^i} + \cdots + a_i \gamma^{2^{(t-1)m} 2^i})$$

From (5), a_i of A is expanded to t multiples of A^* as follows:

$$a_i = a^*_{<\tau^0 2^i>} = a^*_{<\tau^1 2^i>} = \cdots = a^*_{<\tau^{t-1} 2^i>} \tag{7}$$

where $< x >$ denotes the x mod $mt + 1$ operation. Similarly, from (6), we have

$$a_i = a^*_{<2^i>} = a^*_{<2^m 2^i>} = \cdots = a^*_{<2^{(t-1)m} 2^i>} \tag{8}$$

Let us use the following examples to describe (7) and (8).

Example 1. Let $m = 7$ and an element A be represented as $(a_0, a_1, \cdots, a_6)$ in GNB. We can find type-4 for $m = 7$. Based on (3), $\tau = 12$ and $mt + 1 = 29$. A in GNB is transferred to $A^* = (a^*_1, a^*_2, \cdots, a^*_{28})$ in PB according to (5) and (7) as follows:

$$a_0 = a^*_{<12^0 2^0>} = a^*_{<12^1 2^0>} = a^*_{<12^2 2^0>} = a^*_{<12^3 2^0>}$$

$$= a^*_0 = a^*_{12} = a^*_{28} = a^*_{17}$$

where $< 12^0 2^0 >= 1$, $< 12^1 2^0 >= 12$, $< 12^2 2^0 >= 28$, and $< 12^3 2^0 >= 17$.

Similarly, we have the following results.

$$a_1 = a^*_2 = a^*_{24} = a^*_{27} = a^*_5, a_2 = a^*_4 = a^*_{19} = a^*_{25} = a^*_{10}, a_3 = a^*_8 = a^*_9 = a^*_{21} = a^*_{20},$$

$$a_4 = a^*_{16} = a^*_{18} = a^*_{13} = a^*_{11}, a_5 = a^*_3 = a^*_7 = a^*_{26} = a^*_{22}, a_6 = a^*_6 = a^*_{14} = a^*_{23} = a^*_{15}.$$

Example 2. Let $m = 7$, thus type-4 is given. Based on another equation (6) and (8), results are listed as follows:

$$a_i = a^*_{<2^i>} = a^*_{<2^m 2^i>} = \cdots = a^*_{<2^{(t-1)m} 2^i>}$$

$$a_0 = a^*_{<2^0>} = a^*_{<2^7 2^0>} = a^*_{<2^{2\times 7} 2^0>} = a^*_{<2^{3\times 7} 2^0>}$$

$$= a^*_1 = a^*_{12} = a^*_{28} = a^*_{17},$$

where $< 2^0 >= 1$, $< 2^7 2^0 >= 12$, $< 2^{2\times 7} 2^0 >= 28$, and $< 2^{3\times 7} 2^0 >= 17$.

Using (8), we can obtain the following results:

$$a_1 = a^*_2 = a^*_{24} = a^*_{27} = a^*_5, a_2 = a^*_4 = a^*_{19} = a^*_{25} = a^*_{10}, a_3 = a^*_8 = a^*_9 = a^*_{21} = a^*_{20},$$

$$a_4 = a^*_{16} = a^*_{18} = a^*_{13} = a^*_{11}, a_5 = a^*_3 = a^*_7 = a^*_{26} = a^*_{22}, a_6 = a^*_6 = a^*_{14} = a^*_{23} = a^*_{15}.$$

For $m = 7$, we correctly expands a GNB into a PB according to (7) and (8). Examples 1 and 2 show such correct results. But, some m values can not correctly expand a GNB to a PB using (7) and (8). The following examples show such results.

Example 3. Suppose $m = 15$, and therefore any one element A be represented as $(a_0, a_1, \cdots, a_{14})$ in GNB. We can find type-2, and $mt + 1 = 31$ for $m = 15$. If A in GNB can be transferred to $A^* = (a_1^*, a_2^*, \cdots, a_{14}^*)$ according to (5) and (7) as follows.

$$a_0 = a_1^* = a_{30}^*, a_1 = a_2^* = a_{29}^*, a_2 = a_4^* = a_{27}^*, a_3 = a_8^* = a_{23}^*, a_4 = a_{16}^* = a_{15}^*,$$

$$a_5 = a_1^* = a_{30}^*, a_6 = a_2^* = a_{29}^*, a_7 = a_4^* = a_{27}^*, a_8 = a_8^* = a_{23}^*, a_9 = a_{16}^* = a_{15}^*,$$

$$a_{10} = a_1^* = a_{30}^*, a_{11} = a_2^* = a_{29}^*, a_{12} = a_4^* = a_{27}^*, a_{13} = a_8^* = a_{23}^*, a_{14} = a_{16}^* = a_{15}^*.$$

We noted that a_0, a_5, and a_{10} are expanded same coefficients of A^*, a_1^* and a_{30}^*. Another coefficients of A have similar results. Such expanding results are not correct. We use another equation, to check whether $m = 15$ has same expanding results.

Example 4. Let $m = 15$, thus type-2 is given. Based on another equation (6) and (8), results are listed as follows:

$$a_0 = a_1^* = a_{30}^*, a_1 = a_2^* = a_{29}^*, a_2 = a_4^* = a_{27}^*, a_3 = a_8^* = a_{23}^*, a_4 = a_{16}^* = a_{15}^*,$$

$$a_5 = a_1^* = a_{30}^*, a_6 = a_2^* = a_{29}^*, a_7 = a_4^* = a_{27}^*, a_8 = a_8^* = a_{23}^*, a_9 = a_{16}^* = a_{15}^*,$$

$$a_{10} = a_1^* = a_{30}^*, a_{11} = a_2^* = a_{29}^*, a_{12} = a_4^* = a_{27}^*, a_{13} = a_8^* = a_{23}^*, a_{14} = a_{16}^* = a_{15}^*.$$

Computation results also give same wrong results.

Why equations (5) and (7) can not give correct expanding results from a GNB to a PB for some m values?

4 Conclusions

The major advantage of normal basis is its almost cost-free square operation. But, the multiplication in normal basis is very difficult. Therefore, some special classes of normal basis such as optimal normal basis and Gaussian normal basis are employed to overcome this problem. In general, Gaussian normal basis is firstly transferred to polynomial basis. Two equations have been found in the literature to transform any one element in Gaussian normal basis to be represented in polynomial basis. The multiplication in polynomial basis is an easy operation. However, we pointed out that Gaussian normal basis can not be correctly transferred to polynomial basis for some m values. We will solve this problem for the future research.

References

1. Miller, V.S.: Use of elliptic curves in cryptography. In: Williams, H.C. (ed.) CRYPTO 1985. LNCS, vol. 218, pp. 417–426. Springer, Heidelberg (1986)
2. Koblitz, N.: Elliptic curve cryptosystems. Mathematics of Computation **48**, 203–209 (1987)
3. Boneh, D., Franklin, M.K.: Identity-based encryption from the weil pairing. SIAM Journal on Computing **32**(3), 586–615 (2003)
4. Rivest, R.L., Shamir, A., Adleman, L.: A method for obtaining digital signatures and public-key cryptosystems. Communications of the ACM **21**, 120–126 (1978)
5. IEEE Standard 1363–2000: IEEE standard specifications for public-key cryptography (January 2000)
6. ANSI X9.62-2005: Public Key Cryptography for the Financial Services Industry: The Elliptic Curve Digital Signature Algorithm (ECDSA). American National Standards Institute (ANSI) (November 2005)
7. Bartee, T.C., Schneider, D.J.: Computation with finite fields. Information and Computing **6**, 79–98 (1963)
8. Mastrovito, E.D.: VLSI architectures for multiplication over finite field $GF(2^m)$. Applied algebra, algebraic algorithms, and error-correcting codes. In: Mora, T. (ed.) Proc. Sixth Int'l Conf., AAECC-6, Rome, pp. 297–309, July 1988
9. Koç, Ç.K., Sunar, B.: Low-complexity bit-parallel canonical and normal basis multipliers for a class of finite fields. IEEE Trans. Computers **47**(3), 353–356 (1998)
10. Itoh, T., Tsujii, S.: Structure of parallel multipliers for a class of fields $GF(2^m)$. Information and Computation **83**, 21–40 (1989)
11. Lee, C.-Y., Yang, C.-S., Meher, B.K., Meher, P.K., Pan, J.-S.: Low-complexity digit-serial and scalable SPB/GPB multipliers over large binary extension fields using $(b, 2)$-way Karatsuba decomposition. IEEE Trans. Circuits and Systems-I: Regular Papers **61**(11), 3115–3124 (2014)
12. Fan, H., Hasan, M.A.: A new approach to subquadratic space complexity parallel multipliers for extended bina-ry fields. IEEE Trans. Computers **56**(2), 224–233 (2007)
13. Huang, W.-T., Chang, C.H., Chiou, C.W., Tan, S.-Y.: Non-XOR approach for low-cost bit-parallel polynomial basis multiplier over $GF(2^m)$. IET Information Security **5**(3), 152–162 (2011)
14. Xie, J., He, J.J., Meher, P.K.: Low latency systolic Montgomery multiplier for finite field $GF(2^m)$ based on pentanomials. IEEE Trans. VLSI Systems **21**(2), 385–389 (2013)
15. Lee, C.-Y., Meher, P.K., Lee, W.-Y.: Subquadratic space complexity digit-serial multiplier over binary extension fields using Toom-Cook algorithm. In: Proc. of 2014 International Symposium on Integrated Circuits (ISIC), Singapore, pp. 176–179, December 10–12, 2014
16. Berlekamp, E.R.: Bit-serial reed-solomon encoder. IEEE Trans. Inf. Theory **IT–28**, 869–874 (1982)
17. Wu, H., Hasan, M.A., Blake, I.F.: New low-complexity bit-parallel finite field multipliers using weakly dual bases. IEEE Trans. Computers **47**(11), 1223–1234 (1998)
18. Wang, M., Blake, I.F.: Bit serial multiplication in finite fields. SIAM J. Disc. Math. **3**(1), 140–148 (1990)
19. Wang, J.-H., Chang, H.W., Chiou, C.W., Liang, W.-Y.: Low-complexity design of bit-parallel dual basis multiplier over $GF(2^m)$. IET Information Security **6**(4), 324–328 (2012)
20. Hua, Y.Y., Lin, J.-M., Chiou, C.W., Lee, C.-Y., Liu, Y.H.: A novel digit-serial dual basis Karatsuba multiplier over $GF(2^m)$. Journal of Computers **23**(2), 80–94 (2012)

21. Pan, J.-S., Azarderakhsh, R., Kermani, M.M., Lee, C.-Y., Lee, W.-Y., Chiou, C.W., Lin, J.-M.: Low-latency digit-serial systolic double basis multiplier over $GF(2^m)$ using subquadratic Toeplitz matrix-vector product approach. IEEE Trans. Computers **63**(5), 1169–1181 (2014)
22. Massey, J.L., Omura, J.K.: Computational method and apparatus for finite field arithmetic. U.S. Patent Number 4,587,627 (May 1986)
23. Wang, C.C., Troung, T.K., Shao, H.M., Deutsch, L.J., Omura, J.K., Reed, I.S.: VLSI architectures for computing multiplications and inverses in $GF(2^m)$. IEEE Trans. Computers **C–34**(8), 709–717 (1985)
24. Reyhani-Masoleh, A.: Efficient algorithms and architectures for field multiplication using Gaussian normal bases. IEEE Trans. Computers **55**(1), 34–47 (2006)
25. Agnew, G.B., Mullin, R.C., Onyszchuk, I.M., Vanstone, S.A.: An implementation for a fast public-key cryptosystem. Journal of Cryptology **3**, 63–79 (1991)
26. Hasan, M.A., Wang, M.Z., Bhargava, V.K.: A modified Massey-Omura parallel multiplier for a class of finite fields. IEEE Trans. Computers **42**(10), 1278–1280 (1993)
27. Kwon, S.: A low complexity and a low latency bit parallel systolic multiplier over $GF(2^m)$ using an optimal normal basis of type II. In: Proc. of the 16th IEEE Symposium on Computer Arithmetic, Santiago de Compostela, Spain, pp. 196–202, June 15–18, 2003
28. Fan, H., Hasan, M.A.: Subquadratic computational complexity schemes for extended binary field multiplication using optimal normal bases. IEEE Trans. Computers **56**(10), 1435–1437 (2007)
29. Lee, C.-Y., Chiou, C.W.: Scalable Gaussian normal basis multipliers over $GF(2^m)$ using Hankel matrix-vector representation. Journal of Signal Processing Systems for Signal Image and Video Technology **69**(2), 197–211 (2012)
30. Chiou, C.W., Chuang, T.-P., Lin, S.-S., Lee, C.-Y., Lin, J.-M., Yeh, Y.-C.: Palindromic-like representation for Gaussian normal basis multiplier over $GF(2^m)$ with odd type-t. IET Information Security **6**(4), 318–323 (2012)
31. Chiou, C.W., Chang, H.W., Liang, W.-Y., Lee, C.-Y., Lin, J.-M., Yeh, Y.-C.: Low-complexity Gaussian normal basis multiplier over $GF(2^m)$. IET Information Security **6**(4), 310–317 (2012)
32. Azarderakhsh, R., Reyhani-Masoleh, A.: Low-complexity multiplier architectures for single and hybrid-double multiplications in Gaussian normal bases. IEEE Trans. Computers **62**(4), 744–757 (2013)
33. Yang, C.-S., Pan, J.-S., Lee, C.-Y.: Digit-serial GNB multiplier based on TMVP approach over $GF(2^m)$. In: Proc. of 2013 Second International Conference on Robot, Vision and Signal Processing, Kitakyushu, Japan, pp. 123–128, December 10–12, 2013
34. Chiou, C.W., Chang, C.-C., Lee, C.-Y., Hou, T.-W., Lin, J.-M.: Concurrent Error detection and Correction in Gaussian Normal Basis Multiplier over $GF(2^m)$. IEEE Trans. Computers **58**(6), 851–857 (2009)
35. Leone, M.: A new low complexity parallel multiplier for a class of finite fields. In: Koç, Ç.K., Naccache, D., Paar, C. (eds.) CHES 2001. LNCS, vol. 2162, pp. 160–170. Springer, Heidelberg (2001)
36. Ash, D.W., Blake, I.F., Vanstone, S.A.: Low complexity normal bases. Discrete Applied Math. **25**, 191–210 (1989)
37. FIPS 186–2: Digital Signature Standard (DSS). Federal Information Processing Standards Publication 186–2, Nat'l Inst. of Standards and Technology (2000)
38. ISO/IEC 11770–3:2008: Information technology - Security techniques - Key management - Part 3: Mechanisms using asymmetric techniques (2008)

Auto-Scaling Mechanism for Cloud Resource Management Based on Client-Side Turnaround Time

Xiao-Long Liu, Shyan-Ming Yuan, Guo-Heng Luo and Hao-Yu Huang

Abstract Currently, providers of Software as a service (SaaS) can use Infrastructure as a Service (IaaS) to obtain the resources required for serving customers. SaaS providers can save substantially on costs by using resource-management techniques such as auto scaling. However, in most current auto-scaling methods, server-side system information is used for adjusting the amount of resources, which does not allow the overall service performance to be evaluated. In this paper, a novel auto-scaling mechanism is proposed for ensuring the stability of service performance from the client-side of view. In the proposed mechanism, turnaround time monitors are deployed as clients outside the service, and the information collected is used for driving a dynamic auto-scaling operation. A system is also designed to support the proposed auto scaling mechanism. The results of experiments show that using this mechanism, stable service quality can be ensured and, moreover, that a certain amount of quality variation can be handled in order to allow the stability of the service performance to be increased.

Keywords Auto scaling · Cloud computing · Turnaround time · Resource management

1 Introduction

Cloud computing [1] with virtualization technologies has become an important trend in the information technology industry. The introduction of Software as a service (SaaS) [2] has changed the scenario information technology usage. Customers use information services directly through the Internet and no longer have to deploy, manage, and monitor the selected software by themselves. Services are

X.-L. Liu · S.-M. Yuan(✉) · G.-H. Luo · H.-Y. Huang
Department of Computer Science, National Chiao Tung University, Hsinchu, Taiwan, ROC
e-mail: {shallen548,lasifu,berberhuang}@gmail.com, smyuan@cs.nctu.edu.tw

T.T. Zin et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 388,
DOI: 10.1007/978-3-319-23207-2_21

chosen based on considering not only functionality, but also performance, stability, security, and quality. Currently, a service is typically provided accompanied by a Service Level Agreement (SLA), which is a contract that addresses factors that customers care about, such as a guaranteed quality and the specific description of a provided service.

In the age of cloud computing, service providers are not required to build the IT infrastructure. Service providers can allocate the demanded resources rapidly using Infrastructure as a service (IaaS); they can use application programming interfaces such as those provided by Amazon Web Service (AWS) [3] or Google Compute Engine (GCE) [4] in order to create, destroy, and configure, for example, virtual machines (VMs), storage, and load balancers. Using this approach allows service providers to save substantially on cost when compared with building and maintaining their own computer centers; this is because service providers can demand a resource capacity that varies according to workload fluctuations.

Auto scaling [5] is a key technique used for ensuring that the quality of a service fits the SLA and for reducing resource wastage. Auto scaling can be used for automatically increasing or reducing resources when required. Currently, several cloud-management services are available, such as AWS CloudWatch [6], RightScale [7], and Scalr [8], which supply basic auto-scaling functionality. The mechanisms used in most products involve monitoring system information such as CPU Utilization, Disk IN/OUT, and Network IN/OUT on the server side in order to trigger system adjustment under certain conditions. Numerous previous studies have focused on determining the relationship between system information and performance experience for the purpose of helping select the metric and the trigger threshold or for estimating the response time of the end user. However, the genuine experiences of clients cannot be readily understood using the collected server-side information because of certain challenges. First, the computing resource is provided by virtualization technology, and each physical machine runs numerous VM instances concurrently. Thus, the capacity of each VM instance is uncertain and differences exist in the CPU steal time and the IN/OUT wait time that are decided by neighbors on the same physical machine. Moreover, the modern system architecture comprises several distinct services and retrieving the details of the capacity of all components might not be possible, which increases the complexity of the estimation method.

This paper introduces a novel auto-scaling mechanism in which the response time is monitored directly from clients outside a service. In the designed system, one or several monitors are deployed in the client-side of view. The monitors repeatedly send requests for sampling response times over certain durations of service time. A coordinator collects this information for the purpose of driving a rule-based auto-scaling mechanism to decide when the service system must scale out. In this system, the capacity details of each component do not have to be obtained for estimating possible turnaround times. A performance drop caused because of any reason can be detected, and system reaction is based on predefined action in order to provide end users with a stable service quality. To validate the proposed mechanism, a serial test was performed for a file-uploading service and

the results were compared with those obtained using another mechanism [9] in which the request-arrival rate serves as the target metric. In this paper, the influences of distinct parameters of the proposed mechanism are also discussed.

The rest of the context is organized as following: We discuss the related works in Section 2; Section 3 introduces the designed system and proposed auto scaling mechanism; The experimental results are presented in Section 4; Finally, we conclude this paper in Section 5.

2 Related Works

The key concerns of cloud-service providers are minimizing costs and satisfying performance requirements. Most of the current cloud resource-management products support a simple rule-based auto-scaling functionality. This allows service managers to use certain system-utilization metrics as indicators in order to determine the number of instances. However, selecting the metric and the threshold required for promising a service quality that satisfies the SLA, such as guaranteeing a turnaround time, can be challenging. Several approaches have been used in order to attempt to identify the mapping relationships between system utilization and performance. In [10][11][12], the measured capacity of VM instance and the request-arrival rate were used for estimating the response time or the cumulative distributions of the response time on a certain number of VM instances. However, these approaches typically cannot be adapted for use in distinct service architectures. Because a system might comprise numerous dissimilar services, obtaining all resource-capacity details might not be possible.

Another approach is to use a direct metric as the indicator when performing the auto scaling [13][14][15]. One SLA-driven system [14] requires only the setting of a request response time between a load balancer and application servers. The load balancer checks the average response time of each server node and the system allocates a new server node when the average response time of any server node is outside a predefined tolerance range. This approach can be used effectively to guarantee stable server-side performance, but the capacity differences of servers or the dispatch policy of the load balancer might lead to excessive scaling out. Thus, this type of scaling cannot be used for guaranteeing the overall performance of a system.

In the approach used in [9], a test was conducted in order to determine the upper bound of requests per second that had an acceptable turnaround time for clients, and the identified upper bound was used as a base for monitoring the genuine requests per second for the purpose of deciding the amount of resources required for allocating VM instance dynamically over the service time. This is a simple and validated method of auto scaling that allows not only the scale-out timing to be determined, but also enables an estimation of the appropriate amounts of additional resources required. Moreover, this approach can be readily used in systems that feature distinct types of architecture, and the system does not have to be modelled in order to estimate the service quality. This approach can also be effectively applied in schedule-based auto scaling after the workload history is used for preconfiguring the scaling

schedule. However, this approach cannot be used for determining the precise amount of resources required. Previous studies have indicated that the performance of the VM instance provided by IaaS varies [16][17] because in IaaS, virtualization technology is used for providing the resource-supply service. In numerous instances, a single physical machine is shared, and each machine cannot be fully separate from other machines. Distinct numbers of VMs or various jobs running on the physical machine, such as the creation of a new VM instance, might substantially affect the performance of each instance. Thus, the performance of each VM instance is not identical, which means that a limit identified using a specific test cannot fit all VM instances in distinct situations or times.

3 System Design

In order to guarantee a stable quality of a service deployed on a cloud platform, a system is designed in this section to support the proposed auto scaling mechanism. The designed framework can be regarded as a service deployment and management toolset. Fig. 1 shows the architecture of the designed system, which includes three layers, the Cloud Service Provider layer, the Service Management layer, and the Monitor layer.

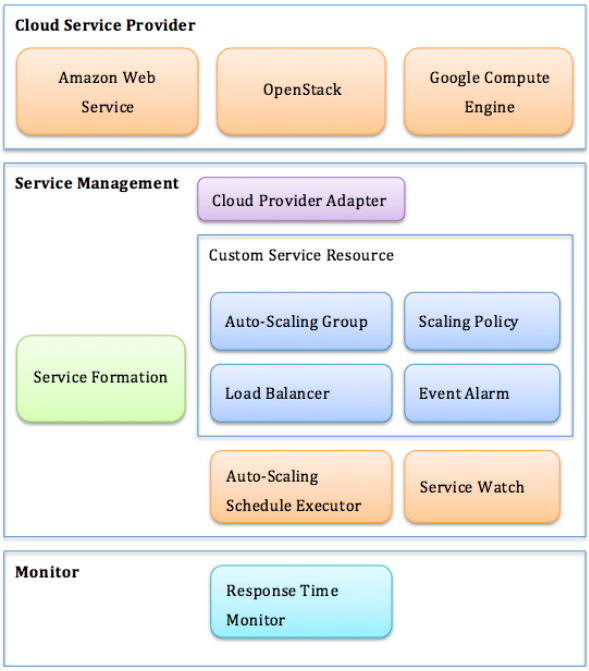


Fig. 1 Architecture of the designed system

3.1 Architecture

The Cloud Service Provider layer supplies the main resources required for running a service, including computing, storage, and networking resources. This layer currently supports AWS, GCE, and other OpenStack-based [18] IaaS systems.

The Service Management layer is the main part of the builder. The flow of operations of this layer mimics the resource-management mechanism of AWS. The Service Formation module can obtain all the resources defined in a configuration file in order to build the architecture of a system automatically. The format of the configuration file is similar to AWS CloudFormation [19], which allows it to be readily applied in extant systems. The module acquires certain resources such as VMs and load balancers through the Cloud Provider Adapter. Furthermore, the module also obtains custom resources that cloud providers do not support. The Service Watch module is a server that collects all the data from various monitors such as the Response Time Monitors in the Monitor layer, and the Event Alarm can use statistical data obtained from the module in order to implement the scaling policies.

The Response Time Monitor repeatedly sends requests to the service in order to evaluate the turnaround time on the client end. The monitor can be installed on numerous local computers to observe the genuine user experience of the performance for the purpose of helping maintain stable service quality.

3.2 Auto Scaling Mechanism

The proposed auto scaling mechanism is used for coordinating the resource provision of a service. Fig. 2 shows an example of this mechanism, the details of which are the following:

1. Response Time Monitors evaluate the response time by using the GET method in order to load a target webpage at regular predefined intervals.
2. The monitors send the response time to the Service Watch module, which collects and classifies these data.
3. Event Alarm repeatedly checks whether specific metrics such as the response time of the GET request are greater than the threshold selected here.
4. If the answer is “true,” trigger the relative-scaling policy and execute it. The answer means that the system might suffer a lack of computing resources, and thus the service must obtain additional resources.
5. All web servers included in the example are organized by Auto Scaling Group, which can use a setting in order to generate numerous identical VMs. The scaling policy increases the capacity of Auto Scaling Group.
6. Auto Scaling Group generates a new web server in the group and allows the service performance to return to the acceptable range.

The flow is continually repeated while the service is online. Moreover, multiple Event Alarms and Scaling Policies can be defined in order to monitor distinct metrics and adjust various resource deployments. The scenario presented in this

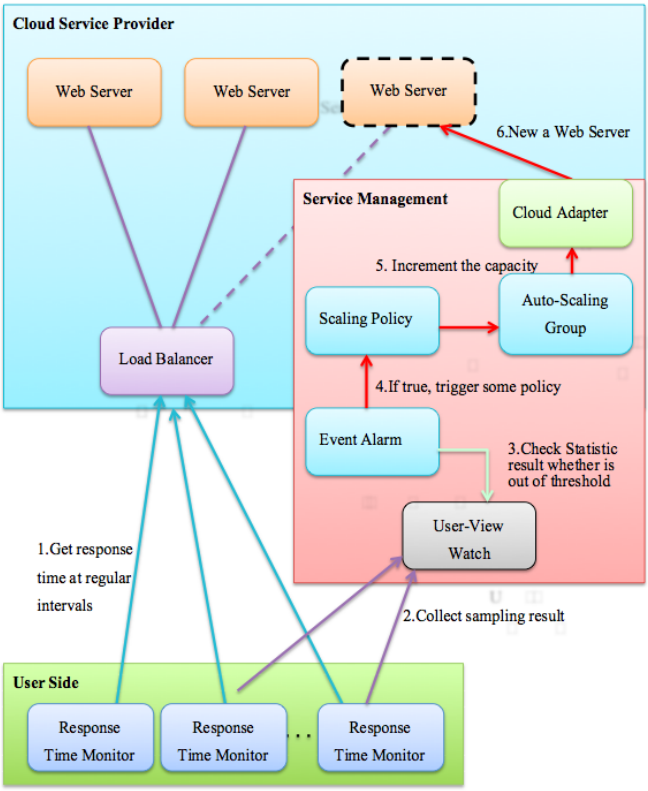


Fig. 2 Example of the proposed auto-scaling mechanism

example is one of a lack of resources. However, the mechanism can also be applied in a situation in which resources are in excess in order to eliminate resources and to save costs.

4 Experimental Results and Evaluation

This section describes the evaluation of the proposed framework in the case of auto-scaling mechanism and the comparison with the approach proposed in [9]. The target system that was tested in the experiment is a file-uploading web service. Each server node deployed a simple receive server, Droopy [20], on a GCE n1-standard-1 type instance. Elastic Load Tester which is based on Locust [21] is used as the test tool to Generate workload. The tester nodes were also built on GCE and used the same instance type because the generated bandwidth was greater than that the testing laboratory used in this study could handle. Each request uploads a 100-KB file to the server through the GCE load balancer. As an indicator of scaling policy, the $\mathrm{AVG} + 3 \times \mathrm{SD}$ of turnaround time was used, where AVG represents the average and SD represents the standard deviation; this is because in a serial-testing display,

approximately 98% of the turnaround times are under this limit. This indicator can be more representative of the overall service quality than the average.

To validate the propose auto scaling mechanism, the following steps were used in the experiment:

- Step 1. Generate a workload configuration in order to simulate a user workload pattern. In each round of the experiment, the same variable workload was used continually for approximately 1 hour to simulate the workload, as shown in Fig. 3.
- Step 2. Generate a 100-KB file as the uploaded file for use in each request.
- Step 3. Prepare a service-deployment configuration file. The initial number of server nodes is one. A single monitor was set on a single instance, and the sampling period set in the experiments was five second. The scale threshold must be set in this step; the thresholds used individually in the experiments were 500, 700, and 1000 ms.
- Step 4. Build a distributed-architecture workload generator featuring four tester nodes.
- Step 5. Start testing. Three rounds of testing were performed using each threshold setting.

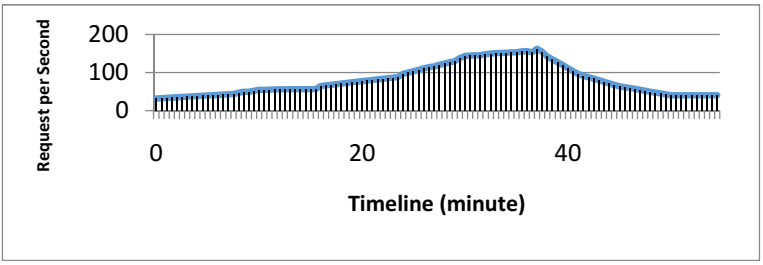


Fig. 3 Workload pattern of simulation

4.1 Results of the Proposed Mechanism

The results in Table 1 show that the average turnaround time measured when the threshold was 500 ms was roughly 15% less than that obtained when the threshold was 1000 ms. Furthermore, the coefficient of variation of the turnaround time at a threshold on 500 ms was approximately 60% of that at a threshold of 1000 ms. The results in Table 2 show that 99% of the requests at a threshold of 500 ms were <300 ms, and that 99% of the requests at a threshold of 1000 ms were <600 ms; however, the instance minute measurement at a threshold of 500 ms was almost 22% more than that at a threshold of 1000 ms. The results show that the use of distinct thresholds can lead to dissimilar overall performance. Setting a small threshold on the turnaround time allowed the stability of the service performance to be enhanced. This result demonstrates that the monitoring method used in the

Table 1 Results of different threshold on auto scaling

Threshold (ms)	500	700	1000
Average of Turnaround time (ms)	108.07	116.74	127.45
SD of Turnaround time (ms)	76.63	93.05	143.83
Coefficient of Variation	70.91%	79.71%	112.84%
Success Request ratio	99.95%	99.92%	99.94%
Instance Minute	137	121	112

Table 2 Cumulative percentage of different threshold on auto scaling

Threshold (ms) / Turnaround Time (ms)	500	700	1000
100	64.54%	50.01%	48.92%
200	98.88%	97.95%	95.81%
300	99.45%	99.08%	98.23%
400	99.62%	99.33%	98.57%
500	99.73%	99.53%	98.93%
600	99.78%	99.60%	99.11%
700	99.80%	99.63%	99.17%
800	99.81%	99.65%	99.19%
900	99.83%	99.67%	99.23%
1000	99.83%	99.69%	99.26%

proposed framework can be employed for ensuring that turnaround times on client-side of view can serve as a dynamic auto-scaling metric.

4.2 Comparison of Different Auto Scaling Mechanisms

The results obtained in this study are then compared with those obtained using the method proposed in [9], where the arrival rate is used as a target metric. For performing the comparison, the arrival-rate threshold in the auto-scaling mechanism proposed by [9] was set as 70 requests per second, and the turnaround-time threshold in our proposed auto-scaling mechanism was set as 700 ms.

As shown in Table 3, the average turnaround time obtained using the approach developed in this study was slightly less than that obtained using the arrival-rate method. The coefficient of variation of the turnaround time in the approach described here was only 70% of that calculated for the comparison method. Moreover, 99% of turnaround time was <300 ms when the Proposed mechanism was used, but 99% of the turnaround time was <600 ms when the arrival-rate approach was used (Table 4). Thus, service quality was more stable when turnaround time from client-side of view was used as a metric than when arrival rate was used as a metric; this indicates that the arrival rate cannot be fully mapped to the turnaround time.

Table 3 Result of different metric on dynamic auto scaling

Item	Proposed mechanism	Arrival Rate mechanism[9]
Average of Turnaround time (ms)	116.74	130.27
SD of Turnaround time (ms)	93.05	141.08
Coefficient of Variation	79.71%	108.30%
Success Request ratio	99.92%	99.95%
Instance Minute	121	110

Table 4 Cumulative percentage of different metric on dynamic auto scaling

Threshold (ms) / Turnaround Time (ms)	Proposed mechanism	Arrival Rate mechanism [9]
100	50.01%	42.58%
200	97.95%	95.47%
300	99.08%	98.14%
400	99.33%	98.54%
500	99.53%	98.89%
600	99.60%	99.09%
700	99.63%	99.15%
800	99.65%	99.19%
900	99.67%	99.24%
1000	99.69%	99.29%

5 Conclusions

This paper proposes and implements a cloud resource-management framework that can be used across multiple cloud platforms in order to deploy, monitor, and scale out a service automatically. The framework setup monitors the service from outside to collect service-performance information for the purpose of driving an auto-scaling mechanism. The proposed framework can be readily applied to certain services that can be used to increase service-system capacity by scaling out in order to meet service-performance requirements and improve user experience. For example, increased numbers of web servers or processing servers can be provided in file-storage services to maintain stable performance during peak times and avoid service interruption in order to save user time and retain user trust in the service.

Acknowledgement This paper was supported by the National Science Council of Taiwan under Grant NSC103-2221-E-009 -133 -MY2.

References

1. Armbrust, M., Stoica, I., Zaharia, M., Fox, A., Griffith, R., Joseph, A.D., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A.: A view of cloud computing. Commun. ACM **53**(4), 50 (2010)
2. Cusumano, M.: Cloud Computing and SaaS As New Computing Platforms. Commun. ACM **53**(4), 27–29 (2010)
3. Amazon Web Services (AWS) - Cloud Computing Services, Amazon Web Services, Inc. http://aws.amazon.com/ (accessed: May 5, 2014)
4. Google Compute Engine - Cloud Computing & infrastructure As A Service, Google COmpute Engine - Cloud Computing & infrastructure As A Service. https://cloud.google.com/products/compute-engine/
5. Lorido-Botrán, T., Miguel-Alonso, J., Lozano, J.A.: Auto-scaling Techniques for Elastic Applications in Cloud Environments, Department of Computer Architecture and Technology, UPV/EHU, EHU-KAT-IK (2012)
6. AWS CloudWatch - Cloud & Network Monitoring Services, Amazon Web Services, Inc. http://aws.amazon.com/cloudwatch/ (accessed: May 5, 2014)
7. RightScale: Cloud Portfolio Management by RightScale, RightScale: Cloud Portfolio Management by RightScale. http://www.rightscale.com
8. Scalr Enterprise Cloud Management Platform. http://www.scalr.com/ (accessed: May 11 2014)
9. Iqbal, W., Dailey, M., Carrera, D.: SLA-Driven Adaptive Resource Management for Web Applications on a Heterogeneous Compute Cloud. In: Jaatun, M.G., Zhao, G., Rong, C. (eds.) Cloud Computing. LNCS, vol. 5931, pp. 243–253. Springer, Heidelberg (2009)
10. Salah, K., Boutaba, R.: Estimating service response time for elastic cloud applications. In: 2012 IEEE 1st International Conference on Cloud Networking (CLOUDNET), pp. 12–16 (2012)
11. Xiong, K., Perros, H.: Service Performance and Analysis in Cloud Computing. In: 2009 World Conference on Services - I, pp. 693–700 (2009)
12. Firdhous, M., Ghazali, O., Hassan, S.: Modeling of cloud system using Erlang formulas. In: 2011 17th Asia-Pacific Conference on Communications (APCC), pp. 411–416 (2011)
13. Chieu, T.C., Mohindra, A., Karve, A.A.: Scalability and Performance of Web Applications in a Compute Cloud. In: 2011 IEEE 8th International Conference on e-Business Engineering (ICEBE), pp. 317–323 (2011)
14. Iqbal, W., Dailey, M., Carrera, D.: SLA-Driven Adaptive Resource Management for Web Applications on a Heterogeneous Compute Cloud. In: Jaatun, M.G., Zhao, G., Rong, C. (eds.) Cloud Computing. LNCS, vol. 5931, pp. 243–253. Springer, Heidelberg (2009)
15. Vasar, M., Srirama, S.N., Dumas, M.: Framework for Monitoring and Testing Web Application Scalability on the Cloud. In: Proceedings of the WICSA/ECSA 2012 Companion Volume, New York, NY, USA, pp. 53–60 (2012)
16. Dejun, J., Pierre, G., Chi, C.-H.: EC2 Performance Analysis for Resource Provisioning of Service-Oriented Applications. In: Dan, A., Gittler, F., Toumani, F. (eds.) ICSOC/ServiceWave 2009. LNCS, vol. 6275, pp. 197–207. Springer, Heidelberg (2010)

17. Lê-Quôc, A., Fiedler, M., Cabanilla, C.: The Top 5 AWS EC2 Performance Problems. DATADOG
18. OpenStack Open Source Cloud Computing Software, OpenStack Open Source Cloud Computing Software. http://www.openstack.org
19. AWS CloudFormation - Configuration Management & Cloud Orchestration, Amazon Web Services, Inc. http://aws.amazon.com/cloudformation/ (accessed: May 5, 2014)
20. stackp/Droopy, GitHub. https://github.com/stackp/Droopy (accessed: May 14, 2014)
21. locustio/locust, GitHub. https://github.com/locustio/locust (accessed: May 5, 2014)

Efficient Digit-Serial Multiplier Employing Karatsuba Algorithm

Shyan-Ming Yuan, Chiou-Yng Lee and Chia-Chen Fan

Abstract This paper presents a efficient digit-serial GF(2^m) multiplier. The proposed architecture using digit-serial of concept to combine the principle of Karatsuba multiplier which can reduce circuit space complexity, also it is suitable for Elliptic Curve Cryptography (ECC) technology. We knows that the password system's operation core is a multiplier, however that password system's multiplier is very big, so it is necessary for reduce the area and time's complexity. This paper is implement three smaller multiplier and digit-serial in FPGA to reduce time and area complexity. This method uses $\frac{3dm}{2}$ AND gate, 6 m + n+ $\frac{3dm}{2}+\frac{m}{2}$ +d-7 XORs and 3 *m*-3 registers. The paper using Altera FPGA Quartus II to simulate four different multipliers, 36 × 36, 84 × 84, 126 × 126 and 204 × 204, and implemented on Cyclone II EP2C70F896C8 experimental platform. The experimental results show that the proposed multipliers have lower time complexity than the existing digit-serial structures. The proposed architecture can reduce the time × space complexity decreasing when the bit-size of multiplier is increasing.

Keywords Karatsuba · Finite field · Digit-serial

1 Introduction

In recent years, the network and mobile phone is very public and important. However, information security is necessary and important. Miler[1] and Koblitz[2] introduction a public key cryptosystem of elliptic curve cryptography (ECC) in 1985. This is widely used in the finite field. First introduction cryptography and proposed a new public gold key cryptographic system which call (elliptic curve

S.-M. Yuan(✉) · C.-Y. Lee · C.-C. Fan
Department of Computer Science, National Chiao Tung University, Hsinchu, Taiwan, ROC
e-mail: smyuan@gmail.com, pp010@mail.lhu.edu.tw, wandy260178@yahoo.com.tw

T.T. Zin et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 388,
DOI: 10.1007/978-3-319-23207-2_22

cryptosystem, ECC). Generally, ECC's mathematics an operation are according GF or GF(2^m), that mean m is finite filed of size; but multiple algorithm is ECC of core operand, so, m is also multiple element of binary digit number. ECC of gold key length far short than other public gold curve cryptosystem, which have lower power with storage capacity smaller necessary of advantage. This characteristic very suitable used in mobile phone, smart card etc.; resource smaller field. Among ECC of core which is used multiplier construction. Multiplier design of bad or not which direct effect to ECC implementation of performance with safety. In the recently, many scholars to work finite filed multiplier of research, which including bit serial, bit parallel and multi bit serial structure, bit parallel multiplier usual adopted Least Significant Bit or Most Significant Bit of way.

For reducing space with time complexity, many researchers proposed some of special polynomial of finite field multiplier. As literature [3, 4] separated all of polynomial, five polynomial with three polynomials utilize matrix of way to develop bit parallel multiplier. As literature is use LSB to implement new three polynomial multipliers which time complexity is $2\sqrt{m}$ pulse circle m is multiplier of bit number. Nevertheless, the others scholar proposed low space complexity bit serial of multiplier. It is only necessary O(m) of space complexity, but opposite time complexity more than length then. For space with time complexity to reach balance, according getting develop different multi bit serial multiplier [8,9] traditional of multi bit multiplier of delaying time is $O(\frac{m}{d})$pulse circle, among d was choice multi bit of bit number size. Literature [10] proposed have bit into the bit structure of multi bit serial multiplier.

Karatsuba-Ofman's(KOA)[11] published 1962, which was first broken situation arrangement integral multiplication sum, due to calculate simple way. So, polynomial version widespread application in GF(2^m) of VLSI multiplier curve cryptosystem. Finite field with KOM of concern thesis succession was proposed by C. Grable [12] designed a 240 bits multiplier implementation in $GF(2^{233})$ above. Literature [13] used Karatsuba algorithm of concept application in finite field $GF(2^m)$ above to propose a low complexity of multiplier.

This paper of others chapters and sections describes as follows: Section 2 describes the mathematical foundations, this thesis will use mathematical concepts. Section 3 contains the algorithm and architecture for the proposed efficient digit-serial multiplier. The experimental results and performance comparison with related work are described in Section 4. Finally, Section 5 is the conclusion of this thesis, and references.

2 Related Works

A. The Finite Field Representation

In this part, we brief introduce finite field of expressing way, finite field GF(2^m) have two 2^m units elements, Each element can express m dimension of vector, moreover each vector all define into it by GF(2), that mean $GF(2^m) = \{A = a_0 + a_1x + a_2x^2 + \cdots + a_{m-1}x^{m-1}\}$, among $a_i = GF(2)$. m must be integer.

In finite field have many basis representation, the most popular bases; polynomial basis (PB), normal basis (NB), dual basis (DB). Finite field of every elements, also can represent $A = a_0 + a_1x + a_2x^2 + \cdots + a_{m-1}x^{m-1}$, among $N = \{a_0, a_1, a_2, \ldots, a_{m-1}\}$, it was finite field of basis of express way. For example: $a_i = x^i$ is polynomial basis, among x is $F(x) = f_0 + f_1x + f_2x^2 + \cdots + f_{m-1}x^{m-1} + x^m = 1 + \sum_{i=1}^{m-1} x^i + x^m$, also that mean $F(x)$=0,

$$x^m = f_0 + f_1x + f_2x^2 + \cdots + f_{m-1}x^{m-1} \tag{1}$$

We can used equation to do multiplier operand of rang stage.

B. Karatsuba Multiplier

In this section, we introduce the Karatsuba application to the original serial structure of polynomial multiplier [16], described in detail as follows:

Element A= $a_0 + a_1x + a_2x^2 + \cdots + a_{m-2}x^{m-2} + a_{m-1}x^{m-1}$ and element B= $b_0 + b_1x + b_2x^2 + \cdots + b_{m-2}x^{m-2} + b_{m-1}x^{m-1}$ is GF(2^m) it is the combination of the irreducible polynomial and the elements A and B are the two lengths of m equations, and element C is the multiplied of elements of AB. We can divide element A and the element B two parts as follows:

$$A = \sum_{i=0}^{m-1} a_ix^i = \sum_{i=\frac{m}{2}}^{m-1} a_ix^i + \sum_{i=0}^{\frac{m}{2}-1} a_ix^i$$

$$= x^{\frac{m}{2}} \sum_{i=0}^{\frac{m}{2}-1} a_{i+\frac{m}{2}}x^i + \sum_{i=0}^{\frac{m}{2}-1} a_ix^i = x^{\frac{m}{2}}A_1 + A_0 \tag{2}$$

$$B = \sum_{i=0}^{m-1} b_ix^i = \sum_{i=0}^{m-1} b_ix^i + \sum_{i=0}^{\frac{m}{2}-1} b_ix^i$$

$$= x^{\frac{m}{2}} \sum_{i=0}^{\frac{m}{2}-1} b_{i+\frac{m}{2}}x^i + \sum_{i=0}^{\frac{m}{2}-1} b_ix^i = x^{\frac{m}{2}}B_1 + B_0 \tag{3}$$

Wherein elements A_0、A_1、B_0 all length of m / 2 bits of the equation, and the product of the elements of A and B of multiplying C can be expressed as follows:

$$C = A_0B_0 + (A_1B_0 + A_0B_1)x^{\frac{m}{2}} + A_1B_1x^m \tag{4}$$

In order to improve the method of calculation of the multiplying of C, we can use equation (6) to improve as the following equation:

$$C = A_0B_0 + ((A_0B_0 + A_1B_1) + (A_1 + A_0)(B_1 + B_0))x^{\frac{m}{2}} + A_1B_1x^m \tag{5}$$

By three kinds of multiplying respectively A_0B_0 、 $(A_0 + A_0)$ $(B_0 + B_1)$ and A_1B_1 multiplication result of the Equation (5). In order to achieve the Equation (5), can use the following three steps to achieve multiplication product:

Step1:

Rate calculation point (KO-EP): three operational points of the elements A, B can be assessed according to the above three multiplications are as follows:

$$KO\text{-}EP\ (A) = (\ A_0\ ,\ A_0 + A_1\ ,\ A_1\ ) \tag{6}$$

$$KO\text{-}EP\ (B) = (\ B_0\ ,\ B_0 + B_1\ ,\ B_1\ ) \tag{7}$$

Step2:
Point to point of computing (KO-PWM): use above of points, and they do it by multiplying the following equation:

$$\begin{aligned} KO\text{-}PWM\ (C) &= KO\text{-}EP\ (A) \times\ KO\text{-}EP\ (B) \\ &= (\ A_0 B_0\ ,(\ A_0 + A_1\ )(\ B_0 + B_1\ ),\ A_1\ B_1\ ) \\ &= (\ t_0\ ,t_1\ ,t_2\ ) \end{aligned} \tag{8}$$

Step 3:
Reconstruction module (KO-R): Finally, based on the operation result of the multiplication, it will t_0 , $t_1$and t_2 , do deoxidize arrangement , the final result of the original multiplication is as follows:

$$\begin{aligned} KO\text{-}EP\ (C) &= (\ \text{KO-PWM}\ (C)) = (\ C_0\ ,C_1\ ,C_2\ ) \\ &= (\ t_0\ ,t_0 + t_1 + t_2\ ,t_2\ ) \end{aligned} \tag{9}$$

According to above three steps, we can draw the following structure Figure 1 , the first step assess computing point, the elements of A and B is divided into two sections, respectively A_0 、 A_1 、 B_0 及 B_1 also use XOR gates respectively A_0 、 $A_1$and B_0 、 $B_1$to add assembly $(\ A_0 + A_1\ )$and $(\ B_0 + B_1\ )$, then the second step will be to obtain three operational point fed three multiplier point -to-point computing, multiplication result t_0, t_1, t_2 , finally, the multiplication results t_0, t_1, t_2 to utilize XOR gate proceeding Karatsuba multiplication reconstruction module and send traditional multiplication results.

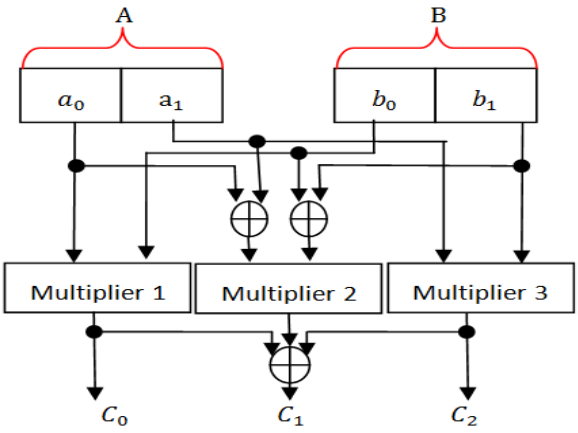


Fig. 1 Traditional of Karatusba multiplier structure

C. Traditional digit-serial multiplier on GF(2^m)

In this parts, we brief introduce multi digit-serial Algorithm. Let GF(2^m) is length m of $F(x)$ not decompose polynomial of composing, For example:

$$A = a_0 + a_1 x + \cdots + a_{m-1} x^{m-1}$$
$$B = b_0 + b_1 x + \cdots + b_{m-1} x^{m-1}$$

Among a and b are $0 \le i \le m$-1,i between 0 and 1, then finite field of element A and B need mod F(x)

$$C=AB \ mod \ F(x) \tag{10}$$

For implementation equation (10), have many different module which can a achieve hardware necessary in finite field of environment. As figure 1 below, we used least significent digit (LSD) multiplier description multi bits serial multiplier construction.

In this thesis will be use element A 、B to separate cutting, more same length of bits operand advance to get result.

If $k = \left\lceil \frac{m}{d} \right\rceil$, among d was selected size, if m is mot dk of multiple which will add 0 in the hightest element.

$$A = (a_0, a_1 \ldots, a_{m-1}, 0 \ldots, 0)$$

Among 0 is $kd - m$ unit bit Element A can express:

$$\mathrm{A}= \sum_{i=0}^{k-1} A_i x^{id}$$

Among $A_0 = a_{id} + a_{id+1} x + \cdots + a_{id+d} x^{d-1}$

Use LSD multiplier module can obtain as below：

$$\begin{aligned} C=&AB \ mod \ F(x) \\ =&\mathrm{B}(A_0 + A_1 x^d + \cdots + A_{k-1} x^{(k-1)d}) \ mod \ F(x) \\ =&C_0 + C_1 + \cdots + C_{k-1} \ \ mod \ F(x) \end{aligned} \tag{11}$$

among

$C_i = B x^{(i)} \ A_i$

$B^{(i)} = B x^{di} \ \ mod \ F(x) = x^d \ \ B^{(i-1)} \ \ mod \ F(x)$

among $0 \le i \le k$-*1*

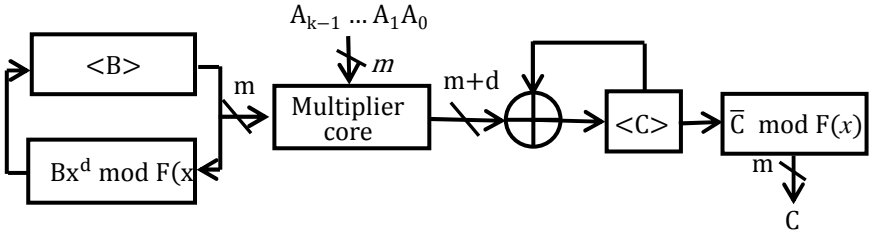


Fig. 2 Traditional LSB digit-serial multiplier

According the equation (11) can draw the LSB digit-serial multiplier in Figure 2. This structure included one multiplier core circuit two register, two reduce polynomial (Bx^d mod F(x) and $\overline{C}$ mod F(x)) and one (m+*d*) piece bit of multiplier. Moreover $\overline{C}$ of register set up zero, according equation (11) of LSD multiplier at $\left\lceil \frac{m}{d} \right\rceil$ piece pulse cycle after, register $\bar{C}$ can obtain $\bar{C}_0 + \bar{C}_1 + \cdots + \bar{C}_{k-1}$, moreover next pulse cycle can implement full complete reduce rank polynomial, among operand obtained $\boldsymbol{C}=\overline{C}\ mod\ F(x)$ ，also, obtained last multiplier operand. And the figure 2 of structure, needed $\left\lceil \frac{m}{d} \right\rceil + 1$ unit of clock.

3 Efficient Digit-Serial Karatsuba Multiplication Algorithm

First description based on Karatsuba two segmentation structure of the law, in to improve the multi-serial multiplication architecture, because the concept of multi-serial, which use once transmission multi-bits of method, also can reduce multiplied time complexity. Karatsuba multiplication of the concept, however, can be a low multiplication space complexity, Karatsuba characteristics computing simple rule, and low computational complexity space, this two concepts combination, can reduce the space complexity and compression time complexity of the following method:

Assume finite field GF (2^m) is irreducible polynomial F (*x*) constructed by polynomial A and B for the finite field GF (2^m) of the two elements, and are as follows:

$$A=A_L + A_H x^{\frac{m}{2}} \qquad B=B_L + B_H x^{\frac{m}{2}}$$

which

$$A_L = a_0 + a_1 + \cdots + a_{\frac{m}{2}-1} x^{\frac{m}{2}-1} \qquad B_L = b_0 + b_1 + \cdots + b_{\frac{m}{2}-1} x^{\frac{m}{2}-1}$$

$$A_H = a_{\frac{m}{2}} + a_{\frac{m}{2}+1} + \cdots + a_{m-1} x^{\frac{m}{2}-1} \qquad B_H = b_{\frac{m}{2}} + b_{\frac{m}{2}+1} + \cdots + b_{m-1} x^{\frac{m}{2}-1}$$

A and B of the product can be expressed as

$$C=(A_L+A_H x^{\frac{m}{2}})(\ B_L+B_H x^{\frac{m}{2}}) \bmod F(x)$$

$$=B_L\left(1+x^{\frac{m}{2}}\right)+A_HB_H\left(x^{\frac{m}{2}}+x^m\right)+(A_LA_H)(B_LB_H)x^{\frac{m}{2}}\ mod\ F(x) \tag{12}$$

According to equation (12), the multiplication contain three sub multiplication A_LB_L、B_L+B_H、$(A_L+A_H)(B_L+B_H)$. Let sub polynomial is

$$A_i=\begin{cases} A_L & for \quad i=0 \\ A_H & for \quad i=1 \\ A_L+A_H & for \quad i=0+1 \end{cases} \qquad B_i=\begin{cases} B_L & for \quad i=0 \\ B_H & for \quad i=1 \\ B_L+B_H & for \quad i=0+1 \end{cases}$$

Then, the product of C can be expressed as

$$C=C_0\left(1+x^{\frac{m}{2}}\right)+C_1\left(x^{\frac{m}{2}}+x^m\right)+C_2x^{\frac{m}{2}}\ mod\ F(x) \tag{13}$$

Which

$$C_i=A_iB_i$$

Suppose d to select the segment length, each polynomial A_i can be expressed as

$$A_i=a_{i,0}+a_{i,1}+\cdots+a_{i,\frac{m}{2}-1}x^{\frac{m}{2}-1}{}_{i,1}x=\sum_{j=0}^{k-1}A_{i,j}x^{jd} \tag{14}$$

which

$$A_{i,j}=\sum_{l=0}^{d-1}a_{i,j}x^l\ ,\ k=\left\lceil\frac{m}{2d}\right\rceil$$

According to equation (13), the sub-product of $C_i=A_iB_i$ can be expressed as

$$C_i=A_iB_i=\sum_{j=0}^{k-1}A_{i,j}B_ix^{jd}$$

$$=(((A_{i,k-1}B_i)x^d+A_{i,k-2}B_i)x^d+\cdots)x^d+A_{i,0}B_i \tag{15}$$

According to (13)-(15), the multiplication algorithm proposed as shown in Figure 4:

According to the Figure 3 of Efficient digit-serial Karatsuba multiplication algorithm, the proposed multiplication architecture can be drawn, as shown in Figure 4, this structure consists of three multiplying core circuit, the three registers, a derating polynomial, three ($m + d$) Each bit adder, three shift circuit, and an XOR gate composition reduction module.

Inputs：$A=A_L+A_H x^{\frac{m}{2}}$ and $B=B_L+B_H x^{\frac{m}{2}}$ are two elements in GF(2^m).

Output：$C=AB \bmod F(x)$.

1. $C_0=0$, $C_1=0$, $C_2=0$.
2. $A_0= A_L$, $A_1= A_H$, $B_0= B_L$, $B_1= B_H$.
3. For $j=k$-1 to 0

*/initialization step

4. $A_{0+1,j}= A_{0,j}$, $+ A_{1,j}$, where $A_{i,j}=(a_{i,dj}, a_{i,dj+1},\cdots, a_{i,dj+d-1})$.
5. $B_{0+1}= B_0+ B_1$.

*/subword product computation step

6. $AB_0= A_{0,j} B_0$.
7. $AB_1= A_{1,j} B_1$.
8. $AB_{0+1}= A_{0+1,j} B_{0+1}$.
9. $C_0= C_0x^{kd}+ AB_0$.
10. $C_1= C_1x^{kd}+ AB_1$.
11. $C_2= C_2x^{kd}+ AB_{0+1}$.
12. endfor

*/ finial polynomial reduction step

13. $C=C_0\left(1+x^{\frac{m}{2}}\right)+C_1\left(x^{\frac{m}{2}}+x^m\right)+C_2x^{\frac{m}{2}} \bmod F(x)$

Fig. 3 Efficient digit-serial Karatsuba multiplication algorithm

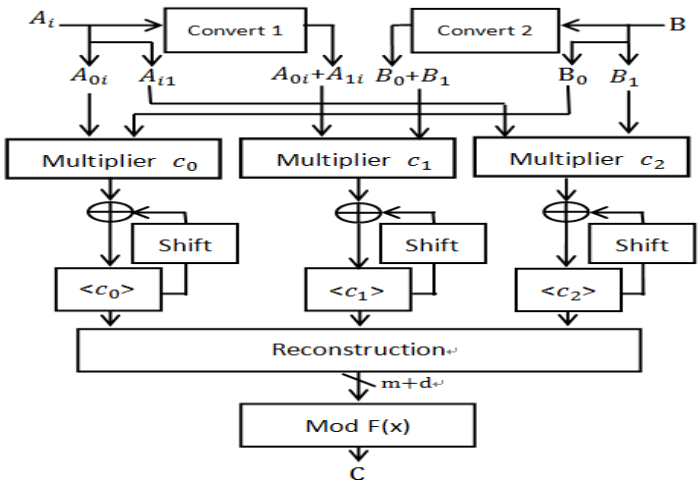


Fig. 4 Proposed multiplier structure

4 Performance Analysis

In this section, we analyze complexity of the time and area on the multiplier hardware. Them in the last chapter, the use of multi-serial architecture combined with the Karatsuba multiplication concept proposed low complexity multi bits multiplier. Structure figure 4, we need to convert 1 and 2, and they need d XOR gates and $\frac{m}{2}$ XOR , then $d \times \frac{m}{2}$ bit multiplier requires three, and three items each time multiplier results sum of XOR gates and c_i register contains a total of 3 ($d \times \frac{m}{2}$) AND gate 3 ($d \times \frac{m}{2}$)XOR gate, (*3m-3*) a temporary devices. Final reduction module requires the XOR gate and mod F(*x*)= $x^m + x^n$ +1 module necessary (*2m+n-3*) units XOR Gate. According to the above statistics, it is possible to obtain in Table 1:

Table 1 Proposed multiplier with present time with space complexity comparison on GF(2^m)

Multiplier	Kumar [13]	Talapatra [15]	Proposed
AND gate	(*m*+2)*d*+2(*d*-1)	*dm*	$\frac{3dm}{2}$
XOR gate	(*m*+1)*d*+2(*d*-1)+(*m*-1)(*n*-1)	*dm*+2*d*	$6m+n+\frac{3dm}{2}+\frac{m}{2}+d-7$
Latch	(*n*+2)*m*+2*d*-(*n*+1)	4*m*+3*d*+1	3*m*-3
Multiplier	0	2*m*	0
Delay time	$\left\lceil \frac{m}{d} \right\rceil + 2$	$2\frac{m}{d}$	$\frac{m}{2d} + 1$

We using Altera FPGA Quartus II on Cyclone II EP2C70F896C8 experimental platform to simulate four different multipliers, 36×36, 84×84, 126×126 and 204×204, and implemented. The Figure 5 and the table 2 are show that the proposed multipliers have lower time × space complexity than 錯誤! 找不到參照來源。, 錯誤! 找不到參照來源。 and 錯誤! 找不到參照來源。. The proposed architecture can reduce the time × space complexity decreasing when the bit-size of multiplier is increasing.

Table 2 The proposed multiplier's reduced the time × space complexity percentage with the propose multiplier

Multiplier / Multiplier size	Kumar [8]	Talapatra[10]	M.Morales [13]
36 bits	80.6%	74.3%	49.6%
84 bits	87.4%	78.0%	66.7%
126 bits	88.3%	77.5%	66.2%
204 bits	91.3%	79.4%	70.4%
Average reduced	86.9%	77.3%	77.3%

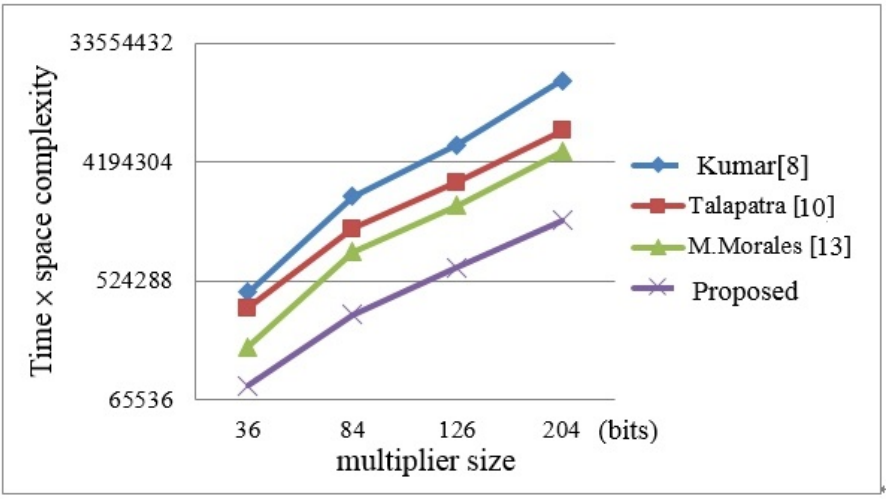


Fig. 5 Comparison the time × space complexity with the propose multiplier and present paper

5 Conclusion

This paper presents a low-complexity multiplier in the limited venue GF (2^m). This method uses the elements cut into the document (s), while for a single use of the Karatsuba the concept to divide into two values, This combined multi byte serial Karatsuba advantage of this approach, which reduce low space complexity but the time complexity does not increase, so, which balance between the time and space complexity, also the multiplier is the greater complexity which savings will come more and more obvious and very suitable for environmental resources is very small but the amount of data required a great environment.

References

1. Miller, V.S.: Use of Elliptic Curves in Cryptography. In: Williams, H.C. (ed.) CRYPTO 1985. LNCS, vol. 218, pp. 417–426. Springer, Heidelberg (1986)
2. Koblitz, N.: Elliptic curve cryptosystems. Mathematics of Computation **48**(177), 203–209 (1987)
3. Lee, C.- Y., Chiou, C.W.: Scalable gaussian normal basis multipliers over GF(2^m) using hankel matrix-Vector representation. J. Signal Processing Systems **69**(2), 197–211 (2012)
4. Fan, H., Hasan, M.A.: A new approach to subquadratic space complexity parallel multipliers for extended binary fields. IEEE Trans. Computers **56**(2), 224–233 (2007)
5. Meher, P.K.: Systolic and non-systolic scalable modular designs of finite field multipliers for reed-solomon codec. IEEE Trans. Very Large Scale Integr. (VLSI) Syst. **17**(6), 747–757 (2009)

6. Xie, J., Meher, P.K., He, J.: Low-complexity multiplier for GF(2^m) based on all-one polynomials. IEEE Trans. Very Large Scale Integr. (VLSI) Syst., **99**, 1–5 (2012)
7. Selimis, G.N., Fournaris, A.P., Michail, H.E., Koufopavlou, O.: Improved throughput bit-serial multiplier for GF(2^m) fields. Integration, the VLSI Journal **42**, 217–226 (2009)
8. Kumar, S., Wollinger, T., Paar, C.: Optimum digit serial GF(2^m) multipliers for curve-based cryptography. IEEE Trans. Computers **55**(10), 1306–1311 (2006)
9. Hariri, A., Reyhani-Masoleh, A.: Digit-Serial Structures for the Shifted Polynomial Basis Multiplication over Binary Extension Fields. In: von zur Gathen, J., Imaña, J.L., Koç, Ç.K. (eds.) WAIFI 2008. LNCS, vol. 5130, pp. 103–116. Springer, Heidelberg (2008)
10. Talapatra, S., Rahaman, H., Mathew, J.: Low complexity digit serial systolic Montgomery multipliers for special class of GF(2^m). IEEE Trans. Very Large Scale Integr. (VLSI) Syst. **18**(5), 487–852 (2010)
11. Karatsuba, A., Ofman, Yu.: Multiplication of multi-digit numbers on automata. Soviet Physics Doklady **7**, 595–596 (1963)
12. Grabbe, C., Bednara, M., Teich, J., von zur Gathen, J., Shokrollahi, J.: FPGA designs of parallel high performance GF(2^m) multipliers. In: Proc. Int. Symp. Circuits Syst. (ISCAS), pp. 268–271 (May. 2003)
13. Ge, Z., Shou, G., Hu, Y., Guo, Z.: Design of Low Complexity GF(2^m) multiplier based on karatsuba algorithm. In: IEEE 13th international Conference on Communication Technology(ICCT), pp. 1018–1022 (September 2011)

Implementation of an FPGA-Based Vision Localization

Wen-Yo Lee, Chen Bo-Jhih, Chieh-Tsai Wu, Ching-Long Shih, Ya-Hui Tsai, Yi-Chih Fan, Chiou-Yng Lee and Ti-Hung Chen

Abstract The robotic version has been widely used in various industry motion control applications, such as object identification, target tracking or environment monitoring, and etc. This paper focuses on studying the real-time FPGA-based implementation of object tracking for a three axes robot. In this work, a unified FPGA implementation for both object identification and target tracking, including basic image processing, image display and target tracking control, is proposed. In addition, target tracking control method with Sobel filter on edge detection, region of interest and motion control. Experimental results show the effectiveness and versatile application ability of the implementation algorithm in target tracking control. Due the flexibility and speed of FPGA hardware, the generated tracking command can be running in very high precision and very high frequency.

Keywords FPGA · Target tracking · Object identification · Robot vision

W.-Y. Lee(✉) · C. Bo-Jhih · C.-Y. Lee · T.-H. Chen
Department of Computer Network and Engineering,
Lunghwa University of Science and Technology, Taoyuan, Taiwan
e-mail: TristanWYLee@mail.lhu.edu.tw

C.-L. Shih
Department of Electrical Engineering,
National Taiwan University of Science and Technology, Taipei, Taiwan

C.-T. Wu
Chang Gung Memorial Hospital, Linkou, Taiwan

Y.-H. Tsai · Y.-C. Fan
Mechanical and System Research Laboratory,
Industrial Technology Research Institute, Hsinchu, Taiwan

T.T. Zin et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 388,
DOI: 10.1007/978-3-319-23207-2_23

1 Introduction

The vision tracking of a manufacturing system, for both pick-place motion control and target tracing, is as important as other aspects of a robot system design, such as fast image processing, high performance object identification and precision motion control. Precision of image localization and speed of the image processing are important facts in design an object tracing control in order to increase manufacturing yield and to reduce production cost and the system settling time.

Due to the advanced development of very large scale integration technology, the field-programmable gate array (FPGA) has been widely used to implement image processing and motion control systems because of its simplicity, programmability, short design cycle, fast time-to-market, low power consumption, and high density. The computing time of an FPGA-based controller can be relatively short regardless of the complexity of the control algorithm because of its parallel processing architecture. A motion controller can be implemented using a single FPGA chip. Therefore, a compact system with low power and a simple circuit is possible. Nowadays, the real-time processing is important in an object tracking system especially in a vision-based motion control system. Chiuchisan have tried to implement a new FPGA-based real-time configurable system for medical image processing [1]. Rodriguez-Araujo, implemented a low cost system-on-chip for localization of UGVs in an indoor iSpace [2]. Chen showed a real-time FPGA-based template matching module for visual inspection for LED defect detection [3]. Hsu gave an idea on FPGA implementation of a real-time face tracking system [4]. Marin discussed about remote programming of network robots within the UJI industrial robotics telelaboratory [5]. Amanatiadis designed a fuzzy area-based image-scaling technique for dynamic neighborhood average image processing [6]. Chinnaiah showed how to implement a shortest path planning algorithm without track using FPGA robot [7]. Ghorbel introduced both hardware and software implementation on FPGA of a robot localization algorithm [8]. Hagiwara, they, used FPGA to research on real-time image processing system for service robots [9]. Saeed showed how to implement the FPGA based real-time target tracking on a mobile platform [10]. Singh based on the Sobel algorithm to implement real-time FPGA based color image edge detection module [11]. Saqui applied the mathematical morphology in object tracking on position-based visual servoing [12].

Because of the prosperous development of advanced vision-based control technology and its simplicity, image processing algorithms and vision servoing are now common used for industry machine motion control applications. Precision positioning machines are required to run with higher speed and higher accuracy. A typical vision servo based on microprocessor or microcontroller is suffering in the speed and precision. Thus the complex programmable logic device, such as field programmable logic device (FPGA), application-specific integrated circuit (ASIC) and system on programmable chip (SoPC), has been stimulating the demand for researcher to develop vision servo capable of very high frequency and very high precision. This paper focuses on studying the FPGA-based vision tracking and target localization robot. In this work, a unified FPGA implementation for both image processing and robot tracking models, including basic image processing model, edge

detection model and object tracking model, is proposed. Experimental results show the effectiveness and versatile application ability of the implementation algorithm in image servoing. Due the flexibility and speed of FPGA hardware, the generated tracking command can be running in very high precision and very high frequency.

The rest of this paper is organized as follows. Section 2 describes system design and image processing methods, basic image processing and target identification. Section 3 presents the FPGA implementation results and experimental results of robot control models. Finally, Section 4 summaries the outcome of the paper and discusses possible implementations for further works.

2 System Design

The experimental set-up of the FPGA-based three axes vision tracking system, as shown in Fig. 1, included the three axes robot manipulator with a CMOS sensor at the end-effector. The Altera DE2-115 Development Board is introduced to implement the tracking system. The three axes robot is built by stepping motors, and the motion commands are generated by pulse analyzer skill which is realized by the Verilog code. A resolution 2,752×2,004 pixels CMOS senor is used to detect the moving target.

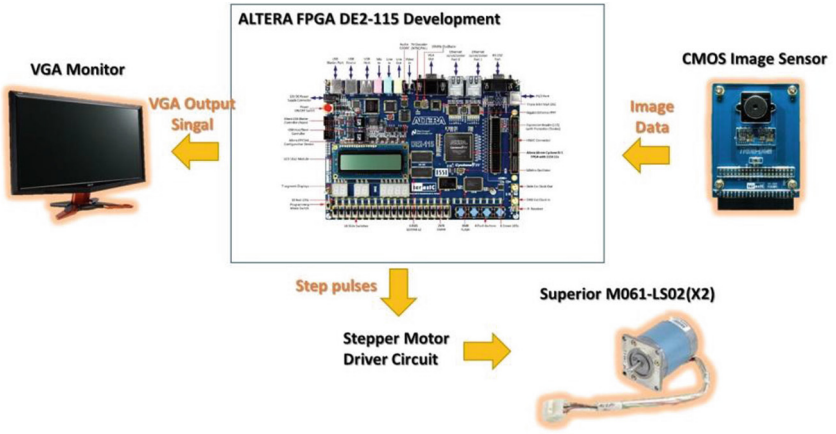


Fig. 1 The FPGA-based three axes vision tracking system.

2.1 The SOC System

The Nios II version 14.1 is introduced to implement the SOC system which composed the basic image process, Sobel edge detection, region of interest, high efficiency memory access, motion control, etc. The SOC-based image process module is shown in Fig. 2.

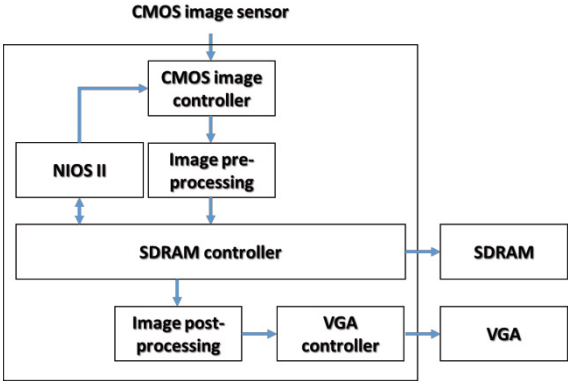


Fig. 2 The SOC-based image process module.

For the moving average filter, the image masking process is solved by LineBuffer module which is a RAM-based shift register. It can perform a high efficiency memory access for the image filtering process. The simulation result of the Linebuffer is shown in Fig. 3. It is similar to the pipeline process, when the clock coms to the 6th clock, then the mask can be calculated simultaneously.

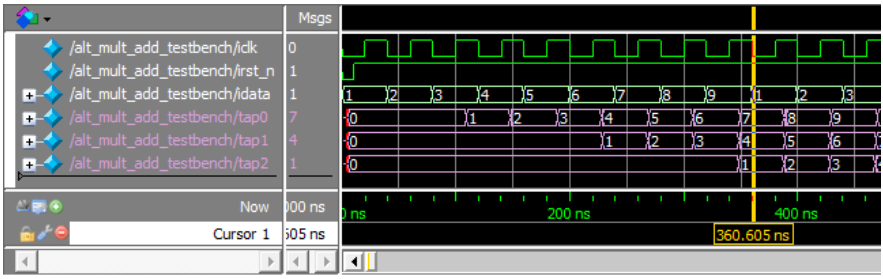


Fig. 3 LineBuffer simulation

The average filter definition is as follows

$$(Y,X) = \left(\sum_{i=-1}^{1}\sum_{j=-1}^{1} M(i,j)P(i+Y,j+X)\right)/9 \ . \tag{1}$$

The major functions of the SOC are CMOS image controller, SDRAM controller for the LineBuffer module, image preprocessing and post processing, and display. The image preprocessing module responses for the basic image processes such that RAW to RGB, Gray image, Sobel edge detection, average filter, region of interest, and centroid, etc. The image data preprocessing block diagram is shown in Fig. 4.

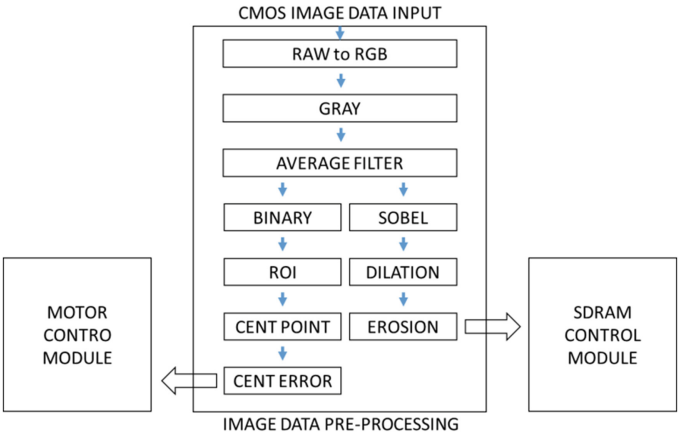


Fig. 4 The image data preprocessing.

There is one important parameter should be mentioned is the ROI. The definition of the region of interest (ROI) is shown by equation (2) and equation (3) which can easily to remove the noise block and get the target on the captured image.

$$xRange\begin{cases}LEFT, & IF(X\times 100)-100<0 \quad then \quad 10\\ RIGHT, & IF(X\times 100)+250>799 \quad then \quad 789\end{cases} \tag{2}$$

$$yRange\begin{cases}UP, & IF(Y\times 100)-100<0 \quad then \quad 10\\ BOTTOM, & IF(Y\times 100)+200>599 \quad then \quad 569\end{cases} \tag{3}$$

2.2 The Target Localization

In a vision-based robot system, the vision sensor gets the image of the target and sends to the image processing center to calculate the target position. The manipulator will get the command to track the target and feedback the target position. This technique is wildly applied on the pick and place application. For demonstration of this idea, we take a post image processing module to rotate and centralize the tracking target. The target image will be shown on the middle of the screen with same orientation of the reference object. The four corners of a rectangular target are taken to calculate the angle of the rotation. It is important to get the long side of the rectangle for rotate the robot vision, since the rotation angle calculation algorithm is based on the information of the long side of it. A simple algorithm is shown in Fig. 5.

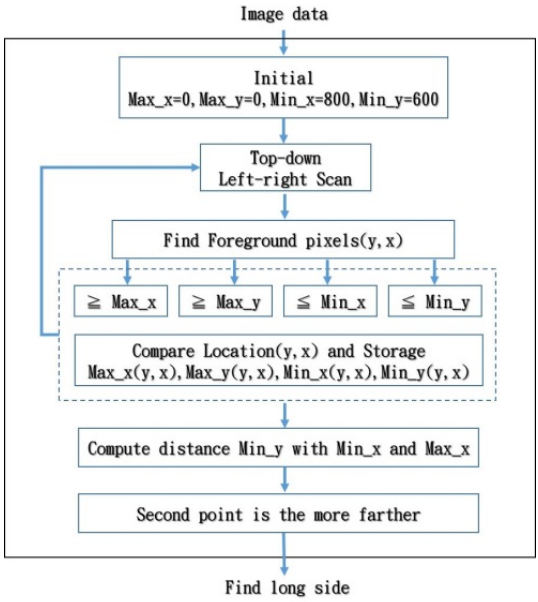
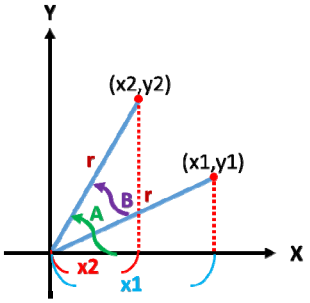


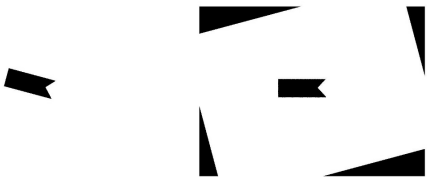
Fig. 5 The algorithm of searching the long side of a rectangle.

According to the Table 1, the vision rotation angle can be find by giving the coordinate (x1,y1) and (x2,y2). The result of the vision rotation experiment is shown in the Fig. 6.

Table 1 The vision rotation formula



$\cos(A) = \frac{x2}{r}$	$\sin(A) = \frac{y2}{r}$
$\cos(A - B) = \frac{x1}{r}$	
$x1 = r \cdot \cos(A - B)$ $x1 = r \cdot \cos(A) \cdot \cos(B) + r \cdot \sin(A) \cdot \sin(B)$ $x1 = x2 \cdot \cos(B) + y2 \cdot \sin(B)$	
$y1 = r \cdot \sin(A - B)$ $y1 = r \cdot \sin(A) \cdot \cos(B) - r \cdot \cos(A) \cdot \sin(B)$ $y1 = -x2 \cdot \sin(B) + y2 \cdot \cos(B)$	



(a) Original image (b) Vision rotate with an angle

Fig. 6 The centralization and rotation result of the target.

3 System Level Design and Experimental Result

The proposed vision system includes three major modules witch are system control module, image process module, and motion control module. It offers a lot of convenient on system design, if the operation system is involved in the original design. The NIOS II is the highest level controller and plays as a user interface. The internal bus delivers the NIOS II comments between the image process module and motion control module. Every module processes in parallel mode and shares data in the SDRAM, which increases the image process efficiency. Since there are individual processes for each module, there is no latency caused by the image process. The system level design block diagram is shown in Fig. 7.

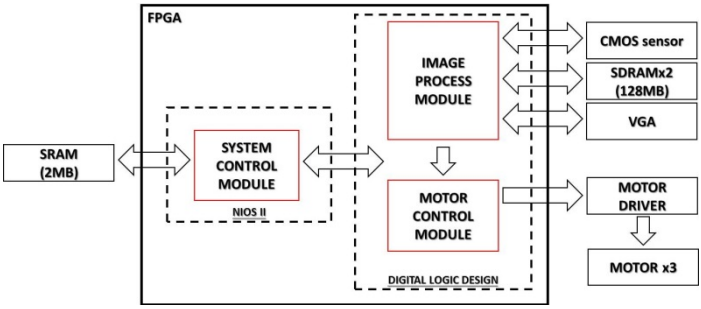


Fig. 7 The system level design block diagram

There are 17 function blocks have been implemented for this study. For speeding up the process time and reducing the latency, the function block is coded by Verilog HDL. The major different from the high level language, such as C++, is that all the function blocks take its own clock to process, so every function block executes their job simultaneously. Fig. 8 shows the system function blocks.

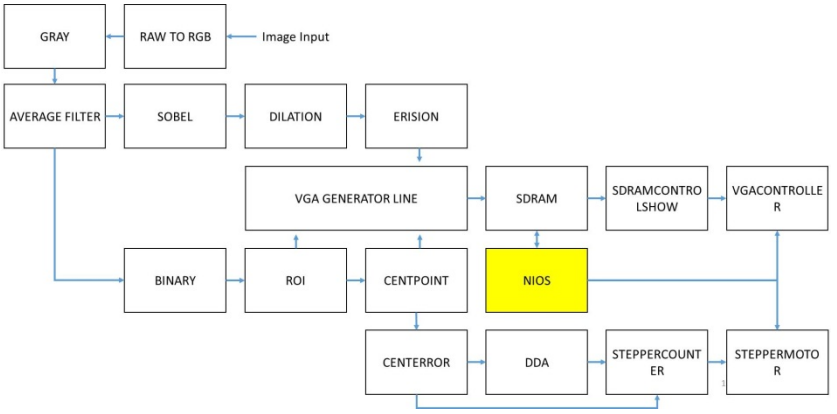


Fig. 8 The system function blocks.

3.1 Experimental Results

A target is put in front of the COMS sensor with a stick holder. Moving the target with the stick the robot will track the target in real-time. The tracking angle of each joint is shown in Fig. 9. The tracking angle is held at a stable point when the target stops moving. The angles from left to right are θ_1, θ_2, and θ_3, respectively.

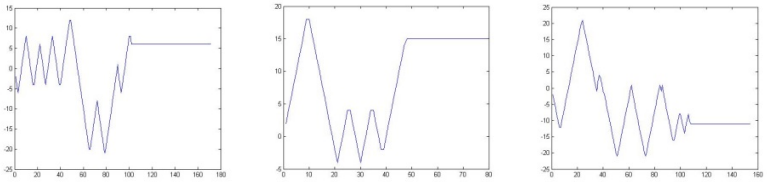
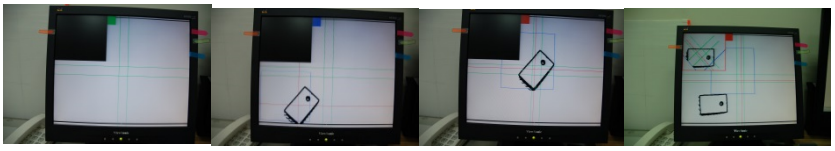


Fig. 9 The tracking angles of each joint.

The experimental result is shown in Fig. 10. It should be mentioned that after the robot locks the target then the robot will rotate it vision angle to let the bottom side of target image parallel to the screen.



(a) Initial state (b) Target move into the vision (c) Lock the target (d) Rotate the vision angle

Fig. 10 Robot tracking result on FPGA

4 Conclusion

This paper proposes a FPGA-based vision tracking algorithms to implement a vision localization robot. The proposed FPGA-based module processes three functions that construct a high speed vision based robot. It is fair to say, the vision based robot is a precision robot, but it integrate the three modules: system control module, image process module, and motion control module which is the fundamental technique for a vision based robot. On the other hand, it offers a whole page of the design skills for how to implement a FPGA-based vision tracking robot, and it can help the researcher to figure out the parallel modules processing in very steps. The proposed implementation method may apply to the industrial markets and remote monitoring markets. The future works are of implementing more function blocks to let the vision tracking robot can be more accuracy. At meantime, we will pay more attention on the message transformation time latency issue on the robot control.

Acknowledgment The work of W.Y Lee is supported by the Taiwan National Science Council (NSC) under Grant NSC102-2221-E-262-018 and Chang Gung Memorial Hospital under Grant CMRPD2C0063.

References

1. Chiuchisan, I.: A new FPGA-based real-time configurable system for medical image processing. In: 2013 IEEE International Conference on E-Health and Bioengineering (EHB), pp. 1–4 (November 2013)
2. Rodriguez-Araujo, J., Rodríguez-Andina, J.J., Fariña, J., Chow, M.Y.: Field-Programmable System-on-Chip for Localization of UGVs in an Indoor iSpace. IEEE Transactions on Industrial Informatics **10**(2), 1033–1043 (2014)
3. Chen, J.Y., Hung, K.F., Lin, H.Y., Chang, Y.C., Hwang, Y.T., Yu, C.K., Chang, Y.J.: Real-time FPGA-based template matching module for visual inspection application. In: 2012 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM), pp. 1072–1076 (July 2012)
4. Hsu, Y.P., Miao, H.C., Tsai, C.C.: FPGA implementation of a real-time image tracking system. In: Proceedings of IEEE SICE Annual Conference 2010, pp. 2878–2884 (August 2010)
5. Marin, R., León, G., Wirz, R., Sales, J., Claver, J.M., Sanz, P.J., Fernández, J.: Remote programming of network robots within the UJI industrial robotics telelaboratory: FPGA vision and SNRP network protocol. IEEE Transactions on Industrial Electronics **56**(12), 4806–4816 (2009)
6. Amanatiadis, A., Andreadis, I., Konstantinidis, K.: Design and implementation of a fuzzy area-based image-scaling technique. IEEE Transactions on Instrumentation and Measurement **57**(8), 1504–1513 (2008)

7. Chinnaiah, M.C., DivyaVani, G., SatyaSavithri, T., Rajeshkumar, P.: Implementation of Shortest path planning algorithm without track using FPGA robot: A new approach. In: 2014 International Conference on Advances in Electrical Engineering (ICAEE), pp. 1–4 (January 2014)
8. Ghorbel, A., Amor, N.B., Jallouli, M., Amouri, L.: A HW/SW implementation on FPGA of a robot localization algorithm. In: Systems, Signals and Devices (SSD), pp. 1-7 (March 2012)
9. Hagiwara, H., Asami, K., Komori, M.: Real-time image processing system by using FPGA for service robots. In: 2012 IEEE 1st Global Conference on Consumer Electronics (GCCE), pp. 720–723 (October 2012)
10. Saeed, A., Amin, A., Saleem, S.: FPGA Based Real-Time Target Tracking on a Mobile Platform. In: 2010 International Conference on Computational Intelligence and Communication Networks (CICN), pp. 560–564 (November 2010)
11. Singh, S., Saini, A.K., Saini, R.: Real-time FPGA Based Implementation of Color Image Edge Detection. International Journal of Image, Graphics and Signal Processing (IJIGSP) 4(12), 19–25 (2012)
12. Saqui, D., Sato, F.C., Kato, E.R., Pedrino, E.C., Tsunaki, R.H.: Mathematical Morphology Applied in Object Tracking on Position-Based Visual Servoing. In: 2013 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 4030–4035 (October 2013)

Supporting Physical Agents in an Interactive e-book

Jim-Min Lin, Che Wun Chiou, Chiou-Yng Lee and Jing-Rui Hsiao

Abstract In recent years, with the advances in information technology and the popularization of computers, obtaining life-around information becomes faster and more convenient. With the emergence of e-books, the reading media is no longer confined to the traditional paper book. People is used to be interactive with lots of interactive media, however most of off-the-shelf e-books offer only data of ordinary flat media, like visual and voice data, and no real touch-interactions there. This study is aimed to use physical agents as interactive media into traditional e-books. Through the performance of the robot body language, users can therefore have profound experience on book context.

1 Introduction

Along with the rapid development of information technology and popularization of all kinds of electronic equipment, obtaining information regarding one's life becomes faster and more convenient. In the past, people gains knowledge or pleasure through the paper books, however, with the emergence of e-books, our reading media is no longer confined to the traditional paper book.

E-book has the features, like small, easy to copy, easy to carry, read elsewhere ... and other features [1]. However, before the e-book reader (e-book device, e-book

J.-M. Lin(✉) · J.-R. Hsiao
Feng Chia University, Taichung 40724, Taiwan
e-mail: jimmy@fcu.edu.tw, pp010@gm.lhu.edu.tw

C.W. Chiou
Chien Hsin University of Science and Technology, Taoyuan County 320, Taiwan
e-mail: cwchiou@uch.edu.tw

C.-Y. Lee
Lunghwa University of Science and Technology, Taoyuan County 333, Taiwan
e-mail: jery6321@gmail.com

T.T. Zin et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 388,
DOI: 10.1007/978-3-319-23207-2_24

reader) come out, most e-books only be developed for the PC environment, and does not have the characteristics that can be read at any time. Therefore the usage of e-books is not high until the mobile device hardware technology becomes mature. Particularly, the world's largest Internet bookstore Amazon e-book reader/carrier "Kindle" makes the development of e-book reader boom. Various manufacturers have sprung up to launch a wide variety of e-book readers/carriers, such as: Sony Reader, ONYX BOOX, Apple iPad ... and so on. The carrier platforms for reading e-books become more diverse, in addition to the existing computers, many appliances and consumer electronics, like PDA, cell phone, e-book readers, TV, watches, refrigerators are likely to become candidate e-book carriers. Through the e-book feature of easy to carry, making it more convenient to get on the huge library resources and enjoy reading any time and any place. Therefore, traditionally, reading convention via the paper media will be gradually converted to reading through the digital media. E-book has more features [2], for example, the characteristics of e-book text, say font size, text color, and display light strength can be adjustable. Therefore it is easy to read in a low light environment. There are some another features. For example, one can read an article by using external speech software. It is really beneficial for the visually impaired persons. In addition, users do not have to manually flip a page as reading an e-book, so there is no risk of damaging a book. Thanks to the rapid development of Internet, it dramatically changes as knowledge is electronically represented and stored on the Internetwork. In such a digital era, the knowledge in the traditional book is then quickly updated to keep up the latest version, which makes a virtually unlimited knowledge source to the e-book readers.

At present, most of the e-book research focused on educational uses. For example, [3] is to get students reading e-books and to explore this way of learning effects on the student's reading ability and attitude. On the other hand, some studies are focused on the commercially-oriented analysis. Through the analysis of major e-book manufacturers and e-book reader product marketing strategies, [4] discusses the development and status of e-book industry. [5] aims to promote users' reading motivation and the learning effectiveness through human/e-book interactions. Up to date, it is rare to apply robots into e-book in literature. To investigate the merits of involving robots in an e-book, we consider to develop a novel 3D human/e-book interaction by adopting real agent (i.e. robot) as a media of human/e-book interactions in the e-book content design. As a result, users can therefore have profound experience on book context through the robot body language performance.

Various interactive media, such as: tone, facial expression, and gestures could be applied in human interactions. However, off-the-shelf e-books usually offer only ordinary flat media, like click linking, picture/video display, and voice playing, but no real touch-interaction there. Studies [6][7] present a variety of tactile vibration device to create the mood in the e-book so that users can have much deeper feelings, and we want to go further to let user/e-book interaction through robots be similar to the human-human communications. Study [8] shows that it is effective to use robots as interactive media in a system's user interface to attract the user's attentions, and thus increases the willingness to continuous use of the system. This is because that robot has a real shape. It not only attracts user of note, but also has a meaningful

way for social interactions. For example, literature [9] proposed that through robot-students interaction, it is increased for students to put their focus on class. Such kind of research also proved that robot is capable of playing a successful social person. Robot's motions not only give user strong impression, but also promote user of interaction interests. In the past, e-book productions need only input text files as the contents of e-books. But, this e-book shows only pure text, and does not have a multimedia content. Today, such kind of e-book no longer meets users' needs. Making multimedia e-books, however, no matter whether it is produced through using software, such as Adobe In Design, Sigil, and so on, a certain degree of program is needed more or less. In order to allow users to easily create multimedia eBook that adopts robot action, this research provides a simple authoring tool allowing users to edit e-books. In our previous works [10][11], we have designed a set of script authoring tools that allow user to edit and to display physical agent's motions. Using this authoring tool, without learning sophisticated computer engineering technology or having art design capabilities, users can easily author physical agent action scripts. Therefore, with our previous research concepts, we will provide an e-book system to facilitate robotic design and authoring tools. Through this tool set, common users can easily produce e-book content integrated with robot motion designs. Users do not need to understand the underlying technology of a robot system.

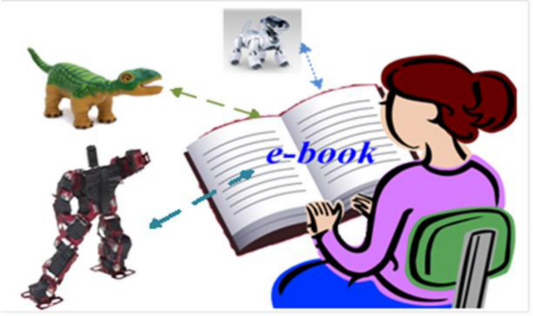


Fig. 1 System conceptual view

Therefore, this study is aimed to propose an application of involving interactive physical agents in e-book systems. We called it a Robot-based e-Book Interaction System (ReBIS). Based on multimedia eBook, in addition to the plain text, image, and voice, the content also involves real robot for users to do real interaction with a robot. In order to be able to verify the practicability of the system, ReBIS is also expected to be applied to fields like education, recreation and so on. In Fig.1, a robot is put next to e-books read by the user as the existence of a contextual entity. A robot may be conceptually representative of a person or an animal in an e-book. Unlike the planar graphs on user's computer screen, robots can show the real presence of an entity in an e-book. When the user is reading an episode in an e-book, he can communicate a physical agent through touching the scene button on the e-book

reader. Robots will then be acting according to user's instructions. Users will read e-books at the same time, and understand the e-book with deeper convey emotion and mood. Future e-books will no longer be just a book. It delivers a new form of human-machine interfaces that will get lots of attentions [12].

2 Related Works

Human computer interaction is the study of the interaction between systems and users. It is often regarded as intersections of computer science, behavioral science, and other fields of studies. The bridge between users and computer's interaction is called Human Computer Interface (HMI), or User Interface (UI) because such an interaction is usually the users who are in control, so the development of UI will need to look after users' needs. The hardware and software instruments include screens that can show characters and objects, keyboards or mouse that users can send messages, or other large-scale computer systems that users can interact with, for example, airplanes and power plants.

2.1 Development of eBook

EBooks are multimedia products that transmit or save words and pictures through electronic ways. The eBooks were applied on encyclopedias or other massive books in early ages, and they can be reduced scale to be only a file with quick search function that users can search for the content quickly. However, some issues may limit the development of eBooks such as resolution of the device and charge capacity of the hardware. The market of eBooks has been growing because of the advancement of smart phones, tablet PCs, the internet speed, and cloud technology. In the past, the eBooks can only be downloaded and then read off-line. Now there are cloud bookshelves and online reading, which the reading process can be synchronized on different devices. Users can read books through popular portable reading devices such as Kindle, iPad, and smart phones. And they get latest books via internet while the contexts will not only be words. Readers can choose what they prefer to read. Some people prefer to read by using cellphones because it's convenient, or to use e-book readers because they will not be tired after a long period time of reading, and others prefer to read on the computers. Different reading habits increase the number of people who use eBooks and make the market grow.

2.2 Formats of E-books

The e-book of early stage is very simple for its text-only version. Later, each manufacturer develops more functions based on their own needs. Therefore, more and more formats, such as the specific format .azw for Kindle, BBeB for Sony Reader, txt, html, pdf, and so on, were developed.

Table 1 Advantages and Disadvantages of Various eBook Formats

	Advantages	Disadvantages
azw	Lots of English e-books high market share Kindle voice function	Only for Kindle (/DX) no Chinese e-books can't compose accurately no Chinese vertical writing
.mobi, .prc	supported by many devices	can't compose accurately no Chinese vertical writing
Lrf (BBeB)	simple files easy to turn pages	only for Sony Reader
.pdf	accurate typesetting Chinese vertical writing	large files hard to turn pages patent protection
.epub	cross-platform, widely used highly interactive for most of e-book publishers for future development	no Chinese vertical writing

In the Table 1, it is obvious that every manufacturer set its format which only supports their own devices. While general formats cannot be too variable, such as .pdf, it can only add text and picture files. Additionally, it consumes lots of system resources when reading it. Lower-tier devices are sometimes slow or accompany with other problems. Furthermore, the various formats bring a big problem for publishers when releasing their e-books. Therefore, the market popularity of e-book has forced manufacturers to create a universal format.

The most common and widely acceptable format is ePub brought up by International Digital Publishing Forum (IDPF). Apart from Kindle, ePub is readable by reader system software on almost all mainstream devices. EPub is supported not only on computers but also on many different mobile devices. Thus, we can do reading anytime and anywhere, and we can search for desired books from online book stores as the Internet is connected. Besides, one more characteristic is that people can make by themselves. Therefore, the ePub definitely will be the mainstream in the future.

3 Development Environment

This research uses the robot Robotinno 1, produced by Innovati. Table 2 is the hardware specification of Robotinno. Three kits are necessary to operate the robot: Servo Commander 16, SYS-214050, and Bluetooth 100M - a Universal Bluetooth Module. Servo Commander 16 is a separate microprocessor module which can control 16 servos, and operate integrated instructions. So that users can directly set up how the servo moves by a fixed speed or common time. There are 250 memory groups for storing servo target positions and moves (speed or time). The SYS-214050 is

Table 2 Robotinno Hardware specifications

Project	Quantity
Servo Commander 16 (BASIC Commander BC1 + Servo Runner A)	1
SYS-214050 (Servo Module)	16
Power adapter (12V, 3A)	1
Nickel-metal hydride battery (10.8V/800mAH)	1
CD with Software	1
Bluetooth 100M	1

the hardware part of robots which support arm rotation angle up to 260 degree. The Bluetooth 100M is the sensing module of robots for communicating with computers.

4 System Structure

4.1 System Architecture and Implementation

Fig.2 shows the system architecture of the proposed eBook system combining the two main modules: Authoring Module and Playback Module. The Playback Module includes two significant sub-modules: Displaying Sub-Module and Robot Motion Sub-Module. Additionally, the Authoring Module includes Multimedia Editing Sub-Module, Action Design Sub-Module, and Integrate Sub-Module. Next, we will introduce the functions and relationships of every individual function block and their implementation details.

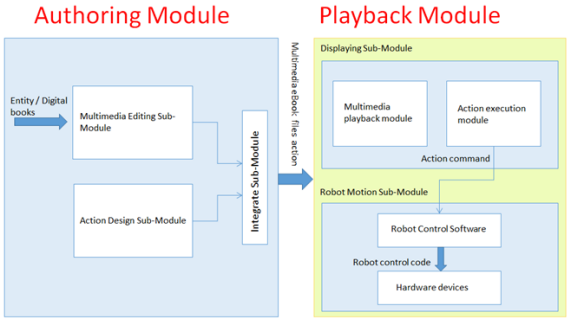


Fig. 2 System architecture

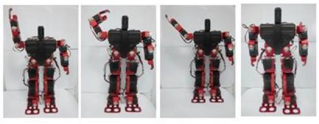
4.2 Multimedia Editing Module

In order to make users to easily use editing eBook, we provide a set of tools that is very easy to get started. Users can make their own eBooks without understanding those complicated applications. Multimedia Editing Sub-Module provides a GUI for not only making various sound and light effects, lift-the-flap scene and type-face…etc., but also directly authoring through importing existing files from various eBook formats, for example those files: .doc, .rtf, .pdf, .txt, epub, as shown in Fig.3a.

4.3 Action Design Module

In the past, if users wanted to design robot actions, they must have a degree of understanding to robots. Thus we provide lots of the picture visualizations of robot basic actions there. Fig.3b is the example of how a user to script a wave action. Besides, through choosing those basic actions, users script a complete sequence action they want. We can see the example in Fig.3c. After scripting those actions, we can name for it and then save. Meanwhile, we can list it on inventory and correlate with the books.

(a) **(b)**



(c)

Fig. 3 (a)Graphical user interface diagram; (b)Picture editing action; (c) Continuous action example

4.4 Integration Sub-module

This sub-module is mainly used to link e-book contents and robot actions together, such that the robot motions can be synchronously played according to the situation described in eBook content. As in Fig.4, we can link the designed robot motions with the selected e-book text. After that, the robot motion design sub-module will generates action instructions for playback of an exclusive operation of the electronic archive in multimedia playback module.

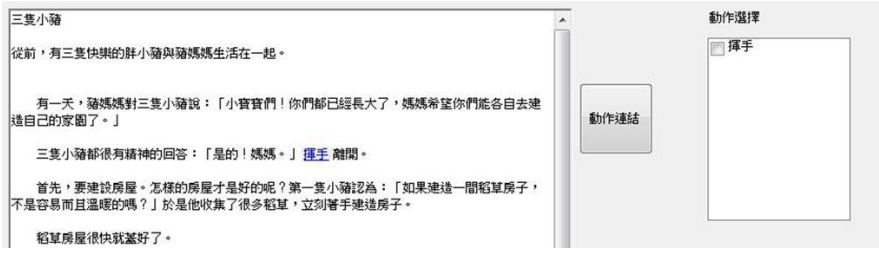


Fig. 4 Robot Motion Linking Interface

4.5 Robot Motion Sub-module

The Robot Motion Sub-Module bears two parts: robot underlying hardware and robot control software. The hardware kit is "Innovati Robotinno 1". Regarding the robot control kit part, it typically utilizes the developing software provided by the manufacturer to design, develop necessary robot control applications. The Robotinno system provides C # programming language to implement its computer-side program because of the need to use Bluetooth to communicate with the robot hardware. The Robotinno manufacturer provides an “innoBASIC Workshop2” for robot control program development. Fig.5 shows the software interface for “innoBASIC Workshop2”.

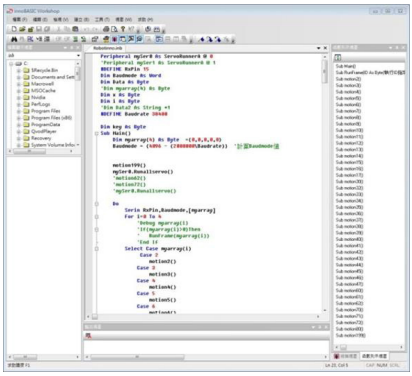


Fig. 5 Software interface for innoBASIC Workshop2

4.6 Displaying Sub-module

The eBook player can read and playback the files, containing multimedia data and robot actions, which we produced in the Multimedia Editing Module. As shown in Fig.6a, the playing screen shows texts, figures, animation, and etc. in the eBook. The blue-color linking text/ figures indicate the paragraph has the situational performance, and users can click the text/ figures to trigger the commands of the robot. The commands will be delivered to robot via Bluetooth and the robot will perform all kinds of the actions matching the situation in the eBook. Through the performance by the robot, users can profoundly understand the situation conveying from the body copy, see Fig.6b.

(a) **(b)**

Fig. 6 (a) System diagram; (b)Playback of robot motions

4.7 Integration Sub-module

Integration sub-module is mainly to make connection between the texts in eBooks and robots actions designed by users, and use robots to convey actions needed in situations of books at the right moment. The relationship between robot motions and texts of books should be synchronized. As shown in Fig.6, we have connected the designed actions in Action Design Sub-Module and the corresponding eBook text. After integrating the robot action commands produced by Multimedia Editing Sub-Module and Action Design Sub-Module, a Multimedia Actions file is generated and sent to Broadcast Sub-Module for delivering to robot driver.

5 Conclusions

In recent years, through the progressive development of information technology, the new ways of human machine interaction appear constantly. Therefore, there are many new styles of human machine interface being developed continually. Then the expectable contribution of this research is to introduce the physical agent into a concept of interactive eBook, and integrate it into an eBook interactive system which

is different from the currently common system, and still reserve the advantages of the original eBook interactive system. The most important issue of the developing human machine interface is how to make general users use any new types of human machine interface without too complicate learning. Hope that the future promotion of this research can make those who no matter members of joining the research or future researchers realize profoundly before designing new types eBook interactive. Meanwhile, the consequence of this research can make the future physical agents become an important reference for the development of eBook.

Acknowledgement This work is partly supported by the Ministry of Science and Technology, Taiwan, Republic of China under the contract of MOST103-2221-E-035-051.

References

1. Gardiner, E., Musto, R.G.: The Electronic Book, The Oxford Companion to the Book, p. 164. Oxford University Press, Oxford (2010)
2. Siegenthaler, E., Wurtz, P., Groner, R.: Improving the Usability of E-Book Readers. Journal of Usability Studies **6**(1), 25–38 (2010)
3. Lin, H.-A.: A study of the influences of e-book reading on tablet PC on the elementary school children's reading ability and attitude, Master's Thesis, Department of Library and Information Science, National Taiwan Normal University, Taiwan (June 2011)
4. Hong, K.: Product and Marketing Strategy Research of eBooks And e-Readers: Based on Amazon, Barnes & Noble, and Apple, Master's Thesis, Department of Business Management, National Taipei University, Taiwan (June 2010)
5. Huang, H.-S.: The Action Research on the Influence of Fourth Graders' Reading Comprehension and Reading Attitude with Digital Reading Instruction, Department of Education, Master's Thesis, National Tainan University, Taiwan (June 2012)
6. Rahman, A.S.M.M., Alam, KM., Saddik, A.E.: A Prototype Haptic E-book System to Support Immersive Remote Reading in a Smart Space. In: IEEE International Workshop on Haptic Audio Visual Environments and Games (HAVE), pp. 61–84 (2011)
7. Alam, K.M., Rahman, A.S.M.M., Saddik, A.E.: HE-Book: A Prototype Haptic Interface for Immersive E-Book Reading. In: World Haptics Conference (WHC), pp. 21–24 (2011)
8. Shinozawa, K., Naya, F., Yamato, J., Kogure, K.: Differences in effect of robot and screen agent recommendations on human decision-making. International Journal of Human-Computer Studies (2004)
9. Kanda, T., et al.: Interactive robots as social partners and peer tutors for children: a field trial. Human-Computer Interaction **19**, 61–84 (2004)
10. Li, K.-Y.: An IDML-Based Robot Control Mechanism, Master's Thesis, Department of information engineering and Computer Science, Feng Chia University (January 2009)
11. You, S.-T.: RDSL: A Domain Specific Language for Robot manipulation, Master's Thesis, Department of information engineering and Computer Science, Feng Chia University (June 2013)
12. Peixuan, C., Huiqing, J.: Interactive electronic story books and instructional design model. ICICE a global Chinese-language e-learning seminar (2003)
13. Zhang, S.-M., et al.: Research on the human computer interaction of E-learning. In: International Conference on Artificial Intelligence and Education (ICAIE) (2010)

A Communication Strategy for Paralleling Grey Wolf Optimizer

Tien-Szu Pan, Thi-Kien Dao, Trong-The Nguyen and Shu-Chuan Chu

Abstract In this paper, a communication strategy for the parallelized Grey Wolf Optimizer is proposed for solving numerical optimization problems. In this proposed method, the population wolves are split into several independent groups based on the original structure of the Grey Wolf Optimizer (GWO), and the proposed communication strategy provides the information flow for the wolves to communicate in different groups. Four benchmark functions are used to test the behavior of convergence, the accuracy, and the speed of the proposed method. According to the experimental results, the proposed communicational strategy increases the speed and accuracy of the GWO on finding the best solution is up to 75% and 45% respectively in comparison with original method.

Keywords Grey wolf optimizer · Parallel grey wolf optimizer · Numerical optimization

1 Introduction

Many optimization problems in engineering, financial, and management fields have been solved successfully by the nature-inspired algorithms. These algorithms have been developed based on the successfully characteristics of biological systems [1]. For example, Genetic algorithms (GAs) were based on the Darwinian evolution of the biological systems [2]. Particle swarm optimization (PSO) was based on the swarm behavior of birds and fish [3]. Artificial bee colony algorithm

T.-S. Pan · T.-K. Dao(✉) · T.-T. Nguyen
Department of Electronics Engineering, National Kaohsiung University of Applied Sciences, Kaohsiung, Taiwan
e-mail: tpan@cc.kuas.edu.tw, jvnkien@gmail.com

S.-C. Chu
School of Computer Science, Engineering and Mathematics, Flinders University, Adelaide, Australia

T.T. Zin et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 388,
DOI: 10.1007/978-3-319-23207-2_25

(ABC) was based on the intelligent foraging behavior of honey bee swarm [4]. Ant colony optimization (ACO) was based on the behavior of ants seeking a path between their colony and a source of food [5, 6]. Cat Swarm Optimization (CSO) was generated by observing the behaviors of cats [7]. Bat algorithm (BA) was based on the echolocation behavior of microbars [8]. Firefly algorithm was based on the flashing light patterns of tropic fireflies [9]. Flower pollination algorithm (FPA) was inspired by the pollination process of flowers [10]. Grey Wolf Optimizer (GWO) was based on the leadership hierarchy and hunting mechanism of gray wolves [11]. All these algorithms have been applied to a wide range of applications.

Based on the algorithms from the nature inspiration, the idea of parallelizing the artificial agents by dividing them into independent subpopulations is introduced into the existing methods such as ant colony system with communication strategies [12], parallel particle swarm optimization algorithm with communication strategies [13], parallel cat swarm optimization [14], Island-model genetic algorithm [15], and parallel genetic algorithm [16]. The parallelized subpopulation of artificial agents increases the accuracy, and extends the global search capacity than the original structure. The parallelization strategies simply share the computation load over several processors. The sum of the computation time for all processors can be reduced, compared with the single processor works on the same optimum problem. In this paper, the concept of parallel processing is applied to the grey wolf optimizer algorithm, and communication strategy for parallel GWO is proposed.

The rest of the paper is organized as follows: a brief review of GWO is given in session 2; our analysis and designs for the parallel GWO is presented in session 3; a series of experimental results and the comparison between original GWO and parallel GWO are discussed in session 4; finally, the conclusion is summarized in session 5.

2 Meta-Heuristic GWO Algorithm

A new meta-heuristic optimization algorithm, namely, Grey Wolf Optimizer (GWO) based on swarm intelligence was proposed in 2014, by Seyedali Mirjalili et al., [11]. It is inspired from observing, imitating, and modeling the leadership hierarchy and hunting mechanism of Grey wolf when searching and attacking for the prey. Grey wolves are considered as apex predators. They mostly prefer to live in a pack, and have a very strict social dominant hierarchy. There are four types of grey wolves in the leadership hierarchy, such as alpha, beta, delta, and omega. So GWO algorithm is guided by the candidate solutions namely: alpha (α), beta (β), delta (δ) and omega (ω). For the type of α is considered the fittest solution, and then β, and δ are considered the second and the third best solutions respectively. Omega (ω) could be assumed the rest of the candidate solutions. Optimizing of GWO algorithm consists of three steps: hunting and searching for prey, encircling prey, and attacking prey. These steps correspond to three constructed mathematical models as follows:

Encircling Prey Mathematical Model: This model is constructed based on social hierarchy of grey wolves. The dominance degree in the leadership hierarchy is formulated in Equations of model as follows:

$$\vec{D} = \left|\overrightarrow{C}.\ \overrightarrow{X_p}(t) - \vec{X}(t)\right| \tag{1}$$

$$\vec{X}(t+1) = \overrightarrow{X_p}(t) - \vec{A} \cdot \vec{D} \tag{2}$$

where $\vec{D}$ is dominance degree, t indicates the current iteration, $\vec{A}$ and $\vec{C}$ are coefficient vectors, $\overrightarrow{X_p}(t)$ is the position vector of the prey, and $\vec{X}(t)$ indicates the position vector of a grey wolf. Equations (1) and (2) are two-dimensional position vector and some of the possible neighbors. The vectors $\vec{A}$ and $\vec{C}$ are calculated as follows:

$$\vec{A} = 2\vec{a} \cdot \overrightarrow{r_1} - \vec{a} \tag{3}$$

$$\vec{C} = 2 \cdot \overrightarrow{r_1} \tag{4}$$

where components $\vec{a}$ are linearly decreased from 2 to 0 over the course of iterations and r_1, r_2 are random vectors in [0, 1]. A grey wolf in the position of (*X, Y*) can update its position according to the position of the prey (*X*, Y**). Different places around the best agent can be reached with respect to the current position by adjusting the value of $\vec{A}$ and $\vec{C}$ vectors. For instance, (*X*–X, Y**) can be reached by setting $\vec{A} = (0,1)$; and $\vec{C} = (1,1)$. The positions are possible to be updated by a grey wolf in 3D space. Note that the random vectors r_1 and r_2 allow wolves to reach any position between the points. The same concept can be extended to a search space with *n* dimensions, and the grey wolves will move in hyper-cubes (or hyper-spheres) around the best solution obtained so far.

Hunting Prey Mathematical Model: The hunting behavior of grey wolves can be simulated when the alpha (best candidate solution), beta, and delta are supposed to have better knowledge about the potential location of prey. Therefore, the first three best solutions are obtained so far and oblige the other search agents (including the omegas) to update their positions according to the position of the best search agents. This simulating model is formulated as follows.

$$\overrightarrow{D_\alpha} = \left|\overrightarrow{C_1}\ \cdot \overrightarrow{X_\alpha}\ - \vec{X}\right|,\ \overrightarrow{D_\beta} = \left|\overrightarrow{C_2}\ \cdot \overrightarrow{X_\beta}\ - \vec{X}\right|, \overrightarrow{D_\delta} = \left|\overrightarrow{C_3}\ \cdot \overrightarrow{X_\delta}\ - \vec{X}\right|, \tag{5}$$

$$\overrightarrow{X_1} = \overrightarrow{X_\alpha} - \overrightarrow{A_1} \cdot \left(\overrightarrow{D_\alpha}\right), \overrightarrow{X_2} = \overrightarrow{X_\beta} - \overrightarrow{A_2} \cdot \left(\overrightarrow{D_\beta}\right), \overrightarrow{X_3} = \overrightarrow{X_\delta} - \overrightarrow{A_3} \cdot \left(\overrightarrow{D_\delta}\right), \tag{6}$$

$$\vec{X}(t+1) = \frac{\overrightarrow{X_1} + \overrightarrow{X_2} + \overrightarrow{X_3}}{3} \tag{7}$$

The position of the prey is estimated by alpha, beta, and delta and other wolves update their positions randomly around the prey during the hunt.

Attacking Mathematical Model: The grey wolves finish the hunt by attacking the prey when it stops moving. This model is determined by variety values of $\vec{A}$. The range of $\vec{A}$ is a random value in the interval [-2a, 2a] and decreased by $\vec{a}$ from 2 to 0 over the course of iterations. The parameter a is to emphasize exploration and exploitation. If the random values of $\vec{A}$ are in [-1, 1], the next position of a search agent can be in any positions between its current position and the position of the prey. The wolves are forced when |A| < 1 to attack towards the prey. Their position is updated based on the dominance of the alpha, beta, and delta. The operators in exploration are to attack towards the prey. When $\vec{A}$ is utilized with random values greater than 1 or less than -1 to oblige the search agent to diverge from the prey, it is known as the exploring mathematical model of divergence. This emphasizes exploration and allows the GWO algorithm to search globally. The grey wolves are forced with |A| > 1 to divergence from the prey to hopefully find a fitter prey. The vector $\vec{C}$ is another component of GWO that favors for exploration. As may be seen in Equation (4), the $\vec{C}$ vector contains random values in [0, 2]. This component provides random weights for prey in order to stochastically emphasize (C > 1) or deemphasize (C < 1) in the effect of prey of defining the distance in Equation (1). The pseudo code of the GWO algorithm is presented in Fig. 1.

Initialize the grey wolf population X_i *(i=1,2 ,..,n)*
Initialize a, A, and C
Calculate the fitness of each search agent
X_α *= the best search agent*
X_β*= the second best search agent*
X_δ*= the third best search agent*
while *(t < Max number of iterations)*
for each search agent
Update the position of the current search agent by Equation (7)
end for
Update a, A, and C
Calculate the fitness of all search agents
Update X_α, X_β and. X_δ
t= t+1
end while
return X_α

Fig. 1 Pseudo code of the GWO algorithm

3 Parallelized GWO with a Communication Strategy

In the parallel structure, several groups are created by dividing the population into subpopulations to construct the parallel processing. Each of the subpopulations evolves independently in regular iterations. They only exchange information

between subpopulations when the communication strategy is triggered. It results to achieve the benefit of cooperation. The parallelized GWO is designed based on original GWO optimization. The wolves in GWO are divided into G subgroups. Each subgroup evolves by GWO optimization independently, i.e. the subgroup has its own wolves as known search agent and finest agents according to the fitness evaluation function. These finest agents among all the wolves in one group will be assigned to the poorer agents in other group, replace them and update agents for each group after running every the fixed iterations. Let G_j be number wolves of the subgroup, where j is the index of the subgroup. While $t \cap R \neq \varphi$, where t is current iteration, and R is the fixed iterations, k agents (where the top k fitness in G_j) will be copied to $G(j+1)$ to replace the same number of agents with the worst fitness, where $j = 0, 1,2,..., G$. The diagram of the parallelized GWO with communication strategy is shown in Figure 2.

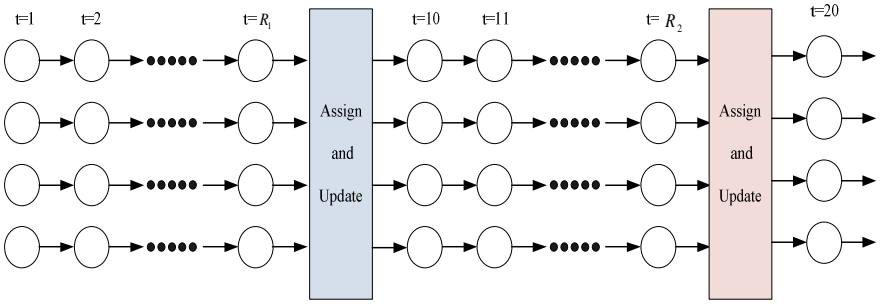


Fig. 2 The diagram of parallel GWO with a communication strategy

1. **Initialization:** Generate wolf population, a, A, and C. Divide them into G subgroups. Each subgroup is initialized by GWO independently. Assign R-the number iterations for executing the communication strategy, N_j the number wolves and X_{ij}^t the solutions for the j-th group, $i = 0, 1,\ldots, Nj-1$; $j = 0, 1,\ldots, G-1$, where G is the number of groups, Nj is the subpopulation size and t is the current iteration and set to 1.

2. **Evaluation:** Evaluate the value of $f(X_{ij}^t)$ for agents in j-th group.

3. **Update: U**pdate the position of the current search agent by Equations (6) and (7), a, A, and C by Equations (3)(4) and (5).

4. **Communication Strategy:** Migrate k best agents among G_j^t to $(j+1)$-th group G_{j+1}^t, mutate G_{j+1}^t by replacing k poorer agents in that group and update every groups in each R iterations.

5. **Termination:** Repeat step 2 to step 5 until the predefined value of the function is achieved or the maximum number of iterations has been reached. Record the best value of the function $f(X_{ij}^t)$ and the best agent among all the wolf position X_{ij}^t.

4 Experimental Results

This section presents simulation results and its comparisons of the parallelized GWO with the original GWO, both in terms of solution quality and speed taken in the number of benchmark function evaluations. Four benchmark functions are used to test the accuracy and the time consumption of the proposed algorithm. All the benchmark functions for the experiments are averaged over different random seeds with 25 runs. The goal of the optimization is to minimize the outcome for all benchmarks. The population size is set the same for all the algorithms in the experiments. The detail of parameter settings of GWO can be found in [11]. The initial range, the dimension and the total iteration number for all test functions are listed in Table 1.

Table 1 The initial range and the total iteration of test standard functions

Test functions	Range	Dim	Iteration
$f_1(x)=\sum_{i=1}^{n-1}[100(x_{1+1}-x_i^2)^2+(x_i-1)^2]$	± 200	30	1000
$f_2(x)=\sum_{i=1}^{n}[x_i^2-10\cos(2\pi x_i)+10]$	± 5	30	1000
$f_3(x)=\frac{1}{4000}\sum_{i=1}^{n}x_i^2-\prod_{i=1}^{n}\cos(\frac{x_i}{\sqrt{i}})+1$	± 500	30	1000
$f_4(x)=[e^{-\sum_{i=1}^{n}(\frac{x_i}{\beta})^{2m}}-2e^{-\sum_{i=1}^{n}x_i^2}]\times\prod_{i=1}^{n}\cos^2 x_i, m=5$	± 20	30	1000

The parameters setting for both parallel GWO and GWO are the initial a, A, and C randomly, the total population size N set to 40, number of group G set to 4, the fixed iteration R set to 10 and the dimension D set to 30. Each function contains the full iterations of 1000 and it is repeated by different random seeds with 25 runs. The final result is obtained by taking the average of the outcomes from all runs. The results are compared with the GWO.

Table 2 compares the quality of performance and time running for numerical problem optimization between parallel GWO and original GWO. It is clearly seen that, almost these cases of testing benchmark functions for parallel GWO are faster convergence than original GWO. It is special case with test function $f_3(x)$, with the mean of value function minimum of total 25 seed runs is 2.21E+00 with average time running equal 1.0080 minutes for parallel GWO evaluation. However, for original GWO these values are 3.21E+00 the mean value function and 3.5150 minutes time running respectively in same executing computer. The performance evaluation for this test function of parallel GWO is up 45% accuracy and 71% time running faster than that in original GWO. The average of four benchmark functions evaluation of minimum function 25 seed runs is 1.42E+05

with average time consuming 0.8777 minutes for parallel GWO and 1.48E+05 with average time consuming 2.8789 for original GWO get 23% accuracy and 66% time speed respectively.

Table 2 The comparison between Parallel GWO and GWO in terms of quality performance evaluation and speed

Test Functions	Performance evaluation		Accuracy %	Time consumption (minutes)		Speed %
	GWO	Parallel GWO	Comparison	GWO	Parallel GWO	Comparison
$f_1(x)$	5.91E+05	5.70E+05	4%	3.2398	0.8053	75%
$f_2(x)$	1.52E+01	1.39E+01	9%	3.3749	0.9274	73%
$f_3(x)$	3.21E+00	2.21E+00	45%	3.5150	1.0080	71%
$f_4(x)$	1.50E-02	1.12E-02	34%	1.3860	0.7700	44%
Average	**1.48E+05**	**1.42E+05**	**23%**	**2.8789**	**0.8777**	**66%**

Figures from 3 to 6 show the experimental results of four benchmark functions in 25 seed runs output with the same iteration of 1000.

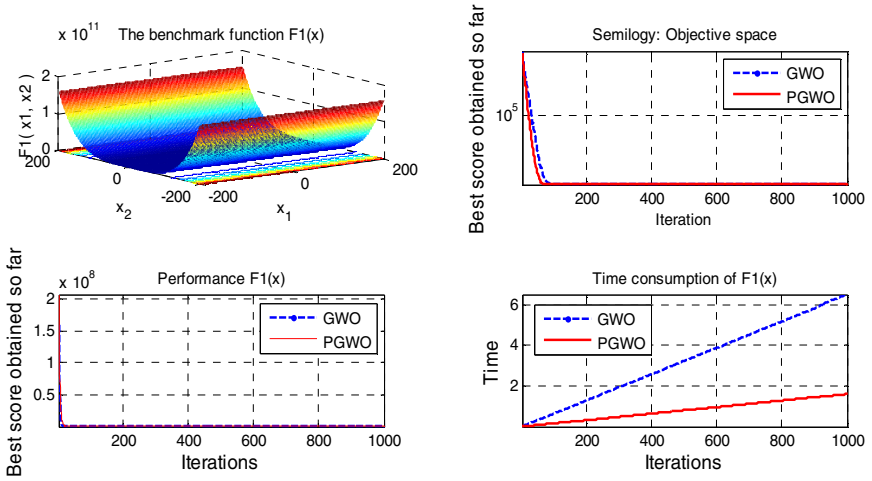


Fig. 3 The experimental results of function F1(x): (a. The bench mark function, b. comparison in semi-logy, c. comparison in performance and d. comparison in time running)

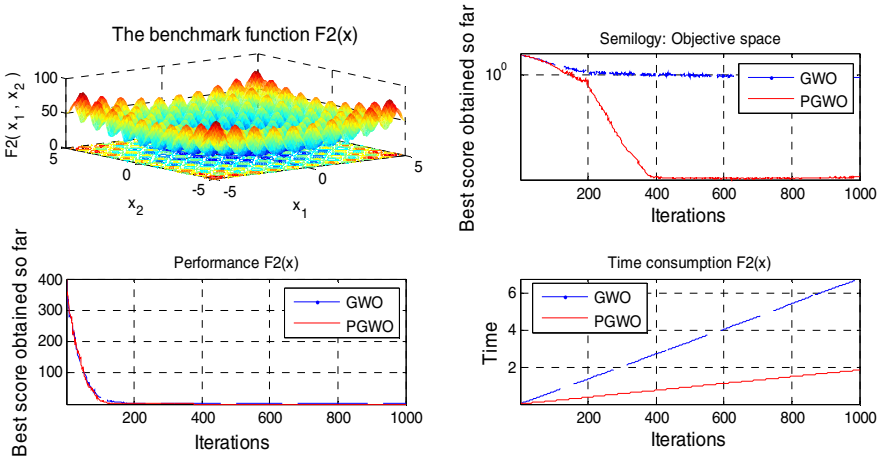


Fig. 4 The experimental results of function F2(x): (a. The bench mark function, b. comparison in semi-logy, c. comparison in performance and d. comparison in time running)

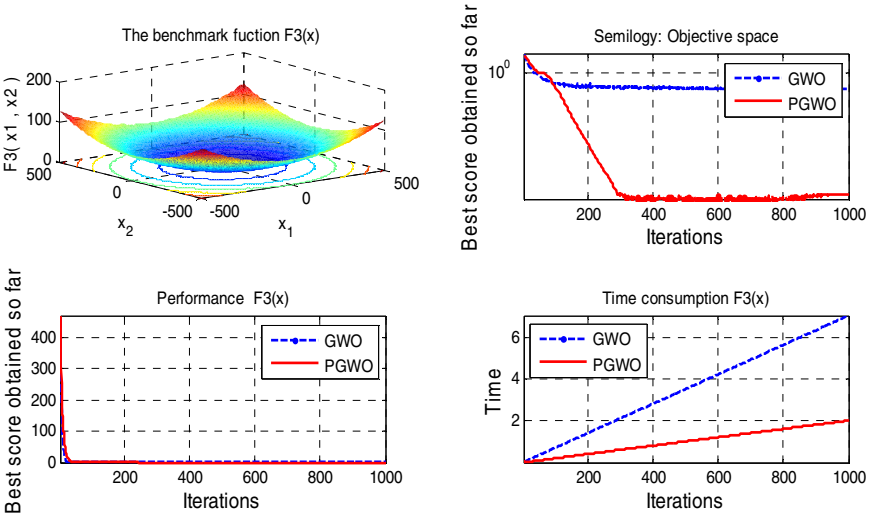


Fig. 5 The experimental results of function F3(x) : (a. The bench mark function, b. comparison in semi-logy, c. comparison in performance and d. comparison in time running)

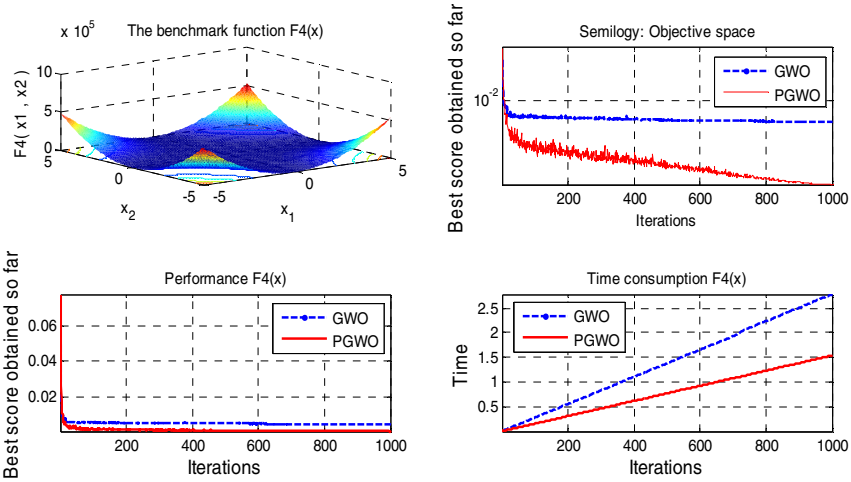


Fig. 6 The experimental results of function F4(x): (a. The bench mark function, b. comparison in semi-logy, c. comparison in performance and d. comparison in time running)

5 Conclusion

This paper, a novel proposed optimization algorithm was presented, namely Parallel GWO. The implementation of parallel optimization algorithms could have important significance for sharing the computation load over several processors, and achieving the cooperation individuals of optimization algorithms. In this new proposed algorithm, the wolves are split into several independent groups based on the original structure of the GWO. The proposed communication strategy provides the information flow for the wolves to assign in different groups. For the communication strategy, the poorer agents in the subgroups will be replaced with new better agents from neighbor subgroups after running each a fixed iteration. This feature is important for the parallel processing devices. The experimental results of the proposed algorithm on a set of various benchmark problems show that the proposed method with communicational strategy increases the accuracy and time running consumption in comparison with the GWO on finding the best solution, such as, the average of testing performance results for above various benchmark problems is 23% accuracy and 66 % time speed respectively.

References

1. Yang, X.-S.: Nature-inspired metaheuristic algorithms. Luniver press (2010)
2. Holland, J.H.: Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence. U. Michigan Press (1975)
3. Kennedy, J., Eberhart, R.: Particle swarm optimization, vol. 4, pp. 1942–1948

4. Karaboga, D.: An idea based on honey bee swarm for numerical optimization, Technical Report-TR06, Erciyes University, Engineering Faculty, Computer Engineering Department, vol. T (2005)
5. Dorigo, M., Caro, G., Gambardella, L.: Ant algorithms for discrete optimization. Artificial Life **5**(2), 137–172 (1999)
6. Dorigo, M., Birattari, M., Stutzle, T.: Ant colony optimization. IEEE Computational Intelligence Magazine **1**(4), 28–39 (2006)
7. Chu, S.-C., Tsai, P.-W.: Computational Intelligence Based on the Behavior of Cats. International Journal of Innovative Computing, Information and Control **3**(1(3)), 8 (2006)
8. Yang, X.-S.: A new metaheuristic bat-inspired algorithm. In: González, J.R., Pelta, D.A., Cruz, C., Terrazas, G., Krasnogor, N. (eds.) NICSO 2010. SCI, vol. 284, pp. 65–74. Springer, Heidelberg (2010)
9. Yang, X.-S.: Firefly algorithm, stochastic test functions and design optimisation. International Journal of Bio-Inspired Computation **2**(2), 78–84 (2010)
10. Yang, X.-S.: Flower pollination algorithm for global optimization. In: Durand-Lose, J., Jonoska, N. (eds.) UCNC 2012. LNCS, vol. 7445, pp. 240–249. Springer, Heidelberg (2012)
11. Mirjalili, S., Mirjalili, S.M., Lewis, A.: Grey wolf optimizer. Advances in Engineering Software **69**, 46–61 (2014)
12. Chu, S.C., Roddick, J.F., Pan, J.-S.: Ant colony system with communication strategies. Information Sciences **167**(1–4), 63–76 (2004)
13. Chu, S.C., Roddick, J.F., Pan, J.-S.: A parallel particle swarm optimization algorithm with communication strategies. Journal of Information Science and Engineering **21**(4), 9 (2005)
14. Tsai, P.-W., Pan, J.-S., Chen, S.-M., Liao, B.-Y., Hao, S.-P.: Parallel Cat Swarm Optimization, pp. 3328–3333
15. Whitley, D., Rana, S., Heckendorn, R.B.: The Island Model Genetic Algorithm: On Separability, Population Size and Convergence. Journal of Computing and Information Technology **1305/1997**, 6 (1998)
16. Abramson, D., Abela, J.: A parallel genetic algorithm for solving the school timetabling problem. In: Proc. of Appeared in 15 Australian Computer Science Conference, no. Hobart, Australia, p. 10 (1991)

Urban Build-Up Building Change Detection Using Morphology Based on GIS

Khaing Cho Moe and Myint Myint Sein

Abstract Rapid urbanization has significant impact on resources and urban environment. In this study, building growth change detection is investigated. To accurate the position of building extraction index, image registration is used that seeks to remove the two-date geometric inconsistent angle with the use of control point selection of latitude and longitude on geographic coordinate system. It is significantly reduce error rates and improve overall accuracy of change detection process. The modified Morphological Building Index (MBI) is applied to extract building features to know how much area has changed. In this system, height information is not considered for building extraction because of without using multispectral band images and Depth. Then, matching-based change rule is applied to obtain changes area of urban region. The experiments show that the proposed method can achieve satisfactory correctness rates with a low level of error rate by comparing with Change Vector Analysis (CVA) method.

Keywords Modified MBI · Control point · Change rule · CVA

1 Introduction

Nowadays, urban sprawl is a worldwide challenge. In a rapidly changing urban environment, remote sensing is a useful technology for change detection and Geographic Information System (GIS) database updating. And Remote Sensing (RS) technology provides data from which updated land cover information can be extracted efficiently and cheaply that is presented in [1]. Extraction of urban area from satellite image is a very important part of GIS features such as updating, geo-referencing and geo spatial data integration.

K.C. Moe(✉) · M.M. Sein
University of Computer Studies, Yangon, Myanmar
e-mail: {khaingchomoe.ucsy,myintucsy}@gmail.com

T.T. Zin et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 388,
DOI: 10.1007/978-3-319-23207-2_26

In particular, if particular objects are of interest, as in our case buildings, it is very difficult to extract those without height information. Many irrelevant changes will be mixed with building changes, when the data are acquired from different sensors or acquired under different imaging geometrics. Moreover, building may be surrounded by dense vegetation; they may have the same color as tree or grass. Awrangjeb *et al*. [2] proposed NDVI only and therefore can't distinguish between a green building and a green tree.

Jin and Davis [3] presented an automated building extraction strategy that simultaneously exploited structural, contextual and spectral information. They applied morphological profiles to extract structural information, reliable contextual information and bright buildings. The final result was obtained by integrating the results of the three different information sources. The problem of this morphological profile causes to commission and omission error. So, X. Hang and L. Zhang [4] and [5] exploited a new method MBI to extract building area without commission and omission error. But this method is applied only on multispectral bands of high resolution satellite images and can't detect low resolution images such as Google Earth image.

Image registration is the most critical operation in remote sensing applications to enable location based referencing and analysis of earth features. Remote sensed data usually contain two types of distortion: radiometric distortion and geometric distortion presented in [6]. There are two major techniques that can be used to correct the various types of geometric distortion. One is to establish correction formulae by modeling the nature of the error sources. Without the knowledge of sensor sources, an image can be registered to a map coordinate system. This is called image-to-map registration in terms of map coordinates (latitudes and longitudes). Using image pixel addresses in terms of a map coordinate base is often referred to as geo-coding.

In this system, we proposed urban growth detection system with three parts: image registration, building feature extraction and change rule. Image registration is carried out by control point with the use of geometric coordinates and then converts it to get image coordinate for the same scene with different angles of yearly images. And then resulted registered image is used to extract building area with modified MBI and finally changed areas are detected by applying change rule. The rest of the paper is organized as follow. Section 2 describes the system overview and section 3 discusses methodology of our proposed system containing image registration, modified MBI and change rule. Section 4 give datasets of using proposed system, section 5 illustrates the experimental results and section 6 concludes the paper.

2 System Overview

The system is divided into three parts. Firstly two input images are rectified by using control point selection of latitude/longitude and geometric coordinate transformation to get registered images for successive ten year. It can accurate

certain position of two images in change detection process as various viewpoints with different sensors. Then, the system is carried on building feature extraction process by modified MBI method [7] and [8] that have human intervention is not required and it is solely unsupervised process. This modified method is effective for building only extraction and convenient in even low resolution satellite imagery such as Google Earth. The challenge of original MBI is that multispectral band of high resolution satellite image can only be applied and haven't considered registration method for better accuracy. Our system can do both high and low resolution satellite image with rectified images. Then resulting building area image is processed by change rule and the final increase buildings are displayed.

3 System Methodology

It contains three parts: Image registration, modified Morphological Building Index and Change rule.

3.1 Image Registration

Image registration is a crucial step in most image processing tasks for which the final result is achieved from the combination of various resources. The ground control point is used to register ten years successive images using image pixel addresses in terms of a map coordinate base. Firstly, the image is grabbed from Google Earth with latitude and longitude coordinate. The area of 0.087 km^2 is for one region with the range of ten second by ten second partitioning of lat/long coordinate. The four ground control points are selected in one region to register successive reference images. Secondly, these obtained lat/long coordinate control points are changed from degree/minute/second format to decimal format. Finally, the coming points are converted by using coordinate conversion method to get XY coordinate. Then, affine transformation is applied to register.

3.1.1 Affine Transformation

The most commonly used registration transformation is the affine transformation which is sufficient to match two images of a scene taken from the same viewing angle but from different position. It can tolerate more complex distortions. Affine transform can be categorized based on the geometric transforms for a planar surface element as translation, rotation, scaling, stretching, and shearing**.**

Obtained control points of image coordinate that is converted from geometric coordinate, building the mapping function and get the affine transformation parameters to resample the sensed image and perform image registration. The general 2D affine transformation can be expressed as shown in the following equation.

$$\begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = \begin{bmatrix} t_x \\ t_y \end{bmatrix} + \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} \quad (1)$$

where (x_2,y_2) is the new transformed coordinate of (x_1,y_1). The matrix $\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$ can be rotation, scale or shear. The scale of both x and y axes can be expressed as

$$Scale = \begin{bmatrix} S_x & 0 \\ 0 & S_y \end{bmatrix} \tag{2}$$

The shear is represented by

$$Shear = \begin{bmatrix} 1 & a \\ 0 & 1 \end{bmatrix}, \quad Shear = \begin{bmatrix} 1 & 0 \\ b & 1 \end{bmatrix} \tag{3}$$

3.2 Modified Morphological Building Index

The basic idea of MBI is to build the relationship between the spectral-structural characteristics of buildings and the morphological operator, which are summarized as follows.

- Brightness
- Local contrast
- Size and Directionality
- Shape

This method uses multispectral bands for high resolution images. Now we use low resolution image of three band color. The modified MBI is defined by describing the characteristic of building feature especially color of building roof and image intensity value. The system runs on low resolution satellite images so their resolution and brightness of intensity values are very low. In order to achieve this problem, modified MBI is proposed as the following steps:

Step 1: Enhancement of Image

The input low resolution registered image is transformed to high contrast image by applying with only red intensity value and stored as the brightness value which is computed by Eq. 4.

$$g = T(f_R(x,y) \tag{4}$$

where $f_R\ (x,y)$ is the intensity transformation of red color-space image, g is the result of enhanced red band image using histogram adjust In [9], Original MBI is applied in multispectral band images of high resolution satellite images. They used enhancement process by using brightness value from this multispectral band. Now our method gives for both high and low resolution of various satellite images with the use of only red color enhancement.

Step 2: Construction of MBI

The spectral-structural characteristics of buildings (e.g., contrast, size and directionality) are represented using the Differential Morphological Profile (DMP).The construction of MBI contains three steps.

(i) White top-hat by Reconstruction can be computed by Eq. 5.

$$W\_TH(d,s) = g - \gamma_b^{re}(d,s) \tag{5}$$

where γ_b^{re} represents the opening-by-reconstruction of the brightness image, and *s* indicates a flat and disk-shaped linear structuring element (SE), respectively.
(ii) Morphological Profiles (MP) of the white top-hat is defined as Eq. 6 and 7.

$$MP_{W\_TH}(s) = W\_TH(s) \tag{6}$$
$$MP_{W\_TH}(s) = 0 \tag{7}$$

(iii) Differential Morphological Profiles (DMP) of the white top-hat is calculated as Eq. 8.

$$DMP_{W\_TH}(s) = \left|MP_{W\_TH}(s+\Delta s) - MP_{W\_TH}(s)\right| \tag{8}$$

where Δs is the interval of the profiles and $s_{min} \le s \le s_{max.}$

MBI is defined as the average of the DMPs of the white top-hat profiles defined in eq. 9 and 10 since buildings have large local contrast within the range of the chosen scales. Thus

$$MBI = \frac{\sum_s DMP_{W\_TH}(s)}{D \times S} \tag{9}$$
$$S = \left(\frac{S_{max} - S_{min}}{\Delta S}\right) + 1 \tag{10}$$

where *D* and *S* denote the numbers of disk and scale of the profiles, respectively.

Step:3 Building extraction
The final building extraction step is decided by using predefined threshold value in order to classify these *MBI (x)* pixels because of different resolutions and image capturing time.

IF MBI(x) $\ge t_1$,
THEN $map_1(x) = 1$
ELSE $map_1(x) = 0$

where *MBI(x)* and $map_1(x)$ indicate the value of *MBI* and the initial label for pixel *x*. t_1 is threshold value and set t_1=5 for the best result for the system.

3.3 Change Rule

After building only areas are extracted by modified MBI in two images, matched-based change rule is applied to get final change/increase building areas.

$$
\begin{aligned}
&If \quad \overline{map}_1(i) \cap \overline{map}_2(i) \\
&\quad then \quad C(i) = 0. \\
&elseif \quad \overline{map}_1(i) \cap map_2(i) \\
&\quad then \quad C(i) = 1. \\
&elseif \quad map_1(i) \cap \overline{map}_2(i) \\
&\quad then \quad C(i) = 0. \\
&else \quad map_1(i) \cap map_2(i) \\
&\quad then \quad C(i) = 0. \\
&end \qquad\qquad\qquad\qquad\qquad\qquad i \in 1,2,3,..., \ N
\end{aligned}
$$

where $map_1(i)$ and $map_2(i)$ are the output value(0 and 1) of modified MBI method. '0' means no building and '1' means building. The i is the same pixel of first and second images and N is the number of pairs of the corresponding building objects where $C(i)$ represents whether the object i is changed, with 0 and 1 for non-change and change, respectively.

4 Datasets

The analysis of the building change detection is carried out based on ten year satellite images from Google Earth acquired between 2004 and 2014. These images include three visible bands such as red, green and blue. The study area lies in the downtown area of Yangon city and covers approximately ten seconds coordinates (0.087 km^2) for one image. It is a typical urban landscape of Myanmar where dense residential and commercial areas are mixed together. Due to the rapid infrastructure construction and updated developing country, the study shows complicated land-cover change.

In order to effectively evaluate building change detection algorithms, we use manually delineated ground truth maps of buildings change. Newly and rebuild building area define changed building ones in our system.

5 Experimental Results

The experiment of the system is tested in Kamaryut Township, Yangon. The change areas of urban region can be successfully extracted using many divisions up to above 1 km^2. The following figure 1 shows user interface of urban change detection system.

Then, if this input area contains four parts, the ground control point must be selected for all parts. These ground control points are latitude and longitude values so geometric transformation is applied to convert XY coordinate point for one image. In four parts, the first image is the range from latitude 16° 49' 44.73''/ longitude 96°07'33.98'' to latitude 16°49'34.73''/ longitude 96°07'43.98''. For image registration, ten year images are set these four control points with their deviation of camera angle.

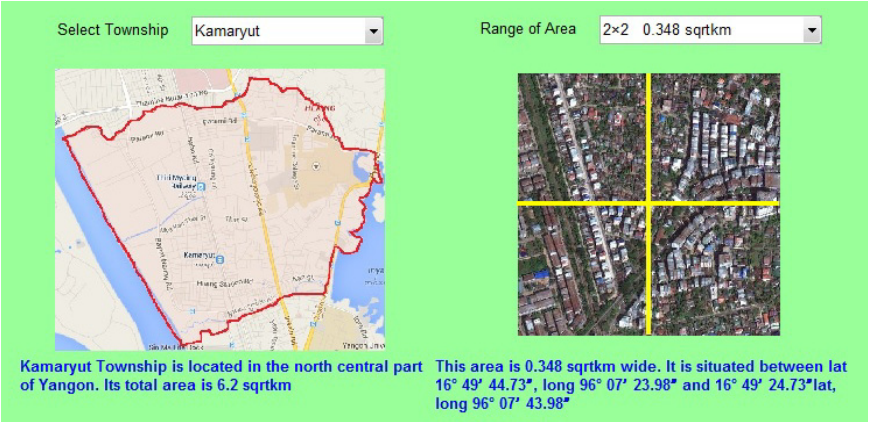


Fig. 1 Input image area located in Kamaryut Township.

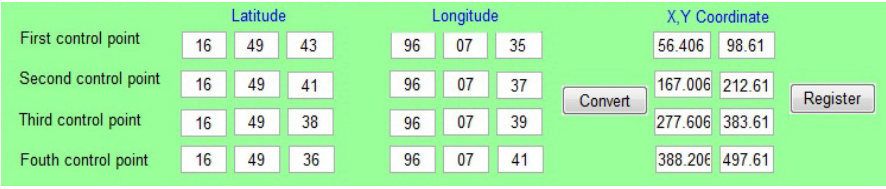
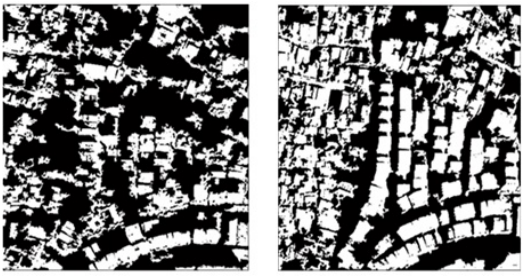


Fig. 2 Geometric transformation of four ground control points for one image.

Fig. 3 Example of one registered image from ten year registered images.

The building extraction process is carried out using modified MBI as shown in figure 4. In these figure, white area means building areas and black area indicates open space areas that have no buildings and can build anyone. The final output is shown in figure 5 in which increase/change building area is detected.

In experiment, errors may occur when many crowded cars on the road lead to miss building extraction because of urban downtown area is our research target. Few building area can't detect in this testing when resolution and contrast is low from Google Earth. The system can detect high resolution commercial satellite images such as Ikonos and can also detect arial images acquired by airplane and high quality camera is used to capture to grab multispectral band image in March, 2014.



(a) 2004 image (b) 2007 image

Fig. 4 Building extraction result using modified MBI.

Fig. 5 Output of the proposed change detection result using change rule

The Panchromatic image can also be used to apply our method that can give greater accureacy and correctness that is from Ikonos satellite images. This is shwon in figure 6.

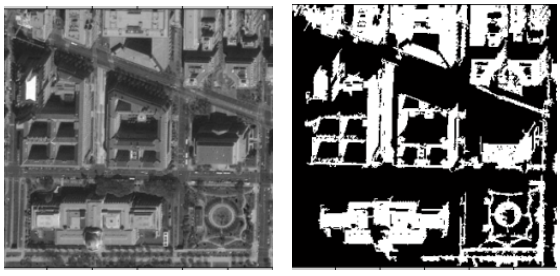


Fig. 6 portion of a 1-meter resolution panchromatic image produced by Ikonos sensor.

In figure 7, arial high resolution satellite image is tested in improved morphology method. The performance of the proposed modified MBI and change rule is compared with change detection approach of Change Vector Analysis (CVA) Method in [10]. CVA is focus on magnitude and angle of different spectral bands for land cover change detection. The magnitude of vectors is calculated by using Euclidean distance. It contains the salt-and-pepper effects of pixel-based change detection approaches.

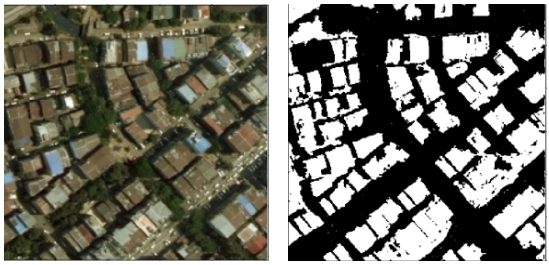


Fig. 7 Arial image of Yangon downtown region.

The evaluation performance is calculated by the following equation. To test the performance of the proposed system, we use these evaluation measures (completeness, correctness, quality) in Table 1.

$$Correctness = \frac{DB}{RB}$$

$$Completeness = \frac{RF}{DB}$$

$$Quality = 100 * \frac{TP}{TP + FN + FP}$$

- TP (True Positive): number of buildings detected both manually and with the automatic approach.
- FP (False Positive): number of buildings detected by the automatic approach but not manually.
- TN (True Negative): number of buildings detected manually but not by the automatic approach.
- FN (False Negative): number of undetected buildings.
- DB: detected buildings at least partially overlapped with the reference buildings.
- RF: reference buildings overlapping the detected buildings.
- The correctness value indicates the percentage of the detected building objects that are at least partially overlapped with the reference buildings.
- The completeness value refers to the percentage of reference buildings overlapped the detected buildings.
- The quality measures the absolute quality of the detection model.

Table 1 Accuracy assessment of the proposed method

Quantitative measures	MBI-based CVA	Proposed Method
Correctness	81.75%	97.87%
Completeness	74.79%	92.45%
Quality	64.09%	89.07%

6 Conclusion

The contribution of this paper is to propose the building growth detection using modified MBI for satellite images which is able to solve various satellite images only with three spectral colors without using multispectral band images because the original MBI is only for multitemporal high-resolution imagery. And the user can know detail latitude and longitude of building in urban area. The characteristic of the proposed method is unsupervised and it is implemented without any training samples so it is able to achieve higher correctness rates and lower average errors than other supervised algorithm. The effectiveness of the proposed method has been validated on Google Earth image of Yangon downtown region in Myanmar with different kinds of building growth including construction, updating and rebuild. The challenge of the proposed system is very low quality satellite image is weak to perform using modified MBI method.

References

1. Tarhan, C.: Detection of environmental and urban change using remote sensing and GIS, thesis book (October 2007)
2. Awrangjeb, M.R., Fraser, C.S.: Automatic building detection using Lidar data multispectral imagery. In: Proceedings of the Digital Image Computing: Techniques and Applications, Sydney, Australia, pp. 45–51 (2010)
3. Jin, X., Davis, C.H.: Automated building extraction from high resolution satellite imagery in urban areas using structural, contextual, and spectral information. EURASIP J. Appl. Signal Process. **14**, 2196–2206 (2005)
4. Hung, X., Zhang, L.: Morphological building/ shadow index for building extraction from high resolution imagery over urban areas. IEEE J. Set. Topic Appl. Earth Obs. Remote Sens. **1**, 161–172 (2012)
5. Hung, X., Zhang, L.: A multidirectional and multiscale morphological index for automatic building extraction from multispectral GeoEye-1 imagery. Photogramm. Eng. Remote Sens. **77**(7), 721–732 (2011)
6. Richard, J.A.: Remote Sensing Digital Image Analysis. Springer (1986)
7. Moe, K.C., Sein, M.M.: An unsupervised technique for building change detection in urban area. Int. J. Computer Application **106**(18), 31–35 (2014)
8. Moe, K.C., Sein, M.M.: Urban growth detection using morphology of satellite image. In: Proceedings of International Conference on Science, Technology, Engineering and Management (ICSTEM 2015), Singapore, February 2015
9. Tang, Y., Huang, X., Zhang, L.: Fault-Tolerant building change detection from urban high-resolution remote rensing imagery. IEEE Geosci. Remote Sens. Lett. **10**(5), September 2013
10. Tong, S., Phan Thi, L., Pham Van, C.: Land cover change analysis using change vector analysis method in Duy Tien District, Ha Nam Province in Vietnam. In: 7th FIG Regional Conference, pp. 19–22, October 2009

Cow Identification by Using Shape Information of Pointed Pattern

Kosuke Sumi, Ikuo Kobayashi and Thi Thi Zin

Abstract Mornitoring cow behavior and performance plays an important role in dairy health and welfare management systems. Due to the increased number of elderly farm workers, the demands for automatic cow monitoring system become a key role. Moreover, the image based technology for a mornitoring system is a promising technique because it is relatively low cost and easy to install. In this aspect, the fundamental and important work to be done is to make an identification of individual cows with high accuracy. Thus in this paper, a simple and effective method for cow identification is introduced by using a modified background subtraction method and histogram based decision process. Specifically, the painted-marks are placed on all black-haired cows and video images are taken. Then the marked region is extracted by using the proposed background subtraction method and histogram based features. Finally, the identification process is performed and some experimental results are shown by using self-collected database taken in the University dairy farm.

1 Introduction

In today dairy farm management research, the potentials of video surveillance technology allow the dairy farmers especially for elderly workers to more closely

K. Sumi(✉)
Department of Electrical and Electronic Engineering,
Graduate School of Engineering, University of Miyazaki, Miyazaki, Japan
e-mail: horizon0125@gmail.com

I. Kobayashi
Field Science Center, Faculty of Agriculture, University of Miyazaki, Miyazaki, Japan

T.T. Zin
Department of Electrical and Systems Engineering, Faculty of Engineering,
University of Miyazaki, 1-1, Gakuen kibanadai-nishi, Miyazaki 889-2192, Japan
e-mail: thithi@cc.miyazaki-u.ac.jp

T.T. Zin et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 388,
DOI: 10.1007/978-3-319-23207-2_27

monitor their herds for identifying and detecting a cow that may have trouble with disease or other functions such as artificial insemination, animal hygiene, food safety and calving. In this aspect, the methodology to identify and recognize individual cow become important for the development of modern dairy farms. To have identification accuracy of milk cow, some approaches by using earmark have been established in Western countries in the early 90s [1]. Also in the late 90s mounting the earmark had become mandatory due to worsening BSE (Bovine Spongiform Encephalopathy) and the issue of foot-and-mouth disease in the late 90s [2]. Since identification accuracy was maintained, we can grasp information of pedigree and artificial insemination in real time by development of individual identification device used earmark and hand-held computer. There, we can rapidly respond to financial management and prevention of epidemic aspects. Independent administrative corporation jointly with domestic animal center had initiated to construct individual identification system for a sizable number of livestock. Recently, standardization of wearing of the earmark has been taken an action of realization processes. However, wearing of the earmark can have side effects for cows such as physical pain and mental anguish. So, in this paper we propose a new image based identification system to identify an individual cow by using video image sequences. The proposed system consists of two processes namely "cow region detection" and "cow identification". We handled it in reference to a study using shape feature in cow identification [3-5].

The rest of the paper is organized as follows. In section 2, we describe the contents and procedure to detect cow region followed by contents and technique to identify cow in section 3. Then in section 4 we present some experimental results to confirm the proposed method. Finally, section 5 concludes the paper by giving remarks and future works.

2 Cow Region Detection

Generally speaking, the classic method of cow region detection utilizes two techniques by using (i) "background difference" and (ii) "inter frame difference". Since a cow has less movement than a human being, we thought that background difference technique is more appropriate than inter frame difference technique. Thus, in this paper we employ the technique of background difference for detecting cow region. The overview of proposed cow region detection algorithm with step by step process as shown in Fig 1.First, as step 1, we perform the preprocessing for the input image. In this step we model an estimated background image, for which rolling of the fixed camera in consecutive image is used in order to remove noise. In step 2, the background image is processed so that it can prevent to recognize a cow as a background. As step 3, we extract a foreground set. In this step, we compute the difference between background image and input image. And, then the region not similar to the background image is extracted as an

output which is typically expressed as a binary image. By using a suitable threshold on background difference value, we make a decision whether the output background difference result is a foreground set. This threshold greatly varies in values due to fix, or be variableness to decide for the background change in each pixel. In step 4, we update an estimated background image by using a recursive median filter because it is difficult to take an early step under the influence of creating every time estimated background. Finally, in step 5, we remove noise by using labeling processing and complement region by using morphology processing. Then the cow region is detected.

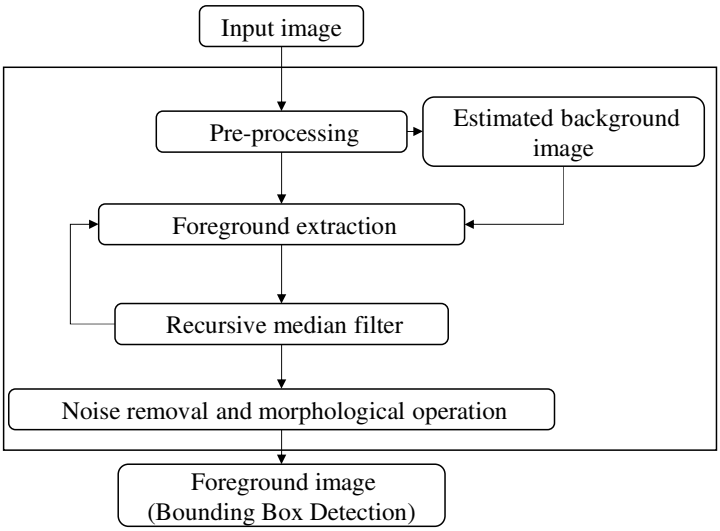


Fig. 1 Overview of cow region detection algorithm

2.1 Pre-processing

2.1.1 Background Image Estimation

Since the cow has lesser movement than a human being, we can make a background estimation for not to be recognizing a cow as a background image. In order to do so for each pixel we extract will be put on top of another and perform a method to substitute a level as an estimated background image. This process is illustrated as shown in Fig 2.

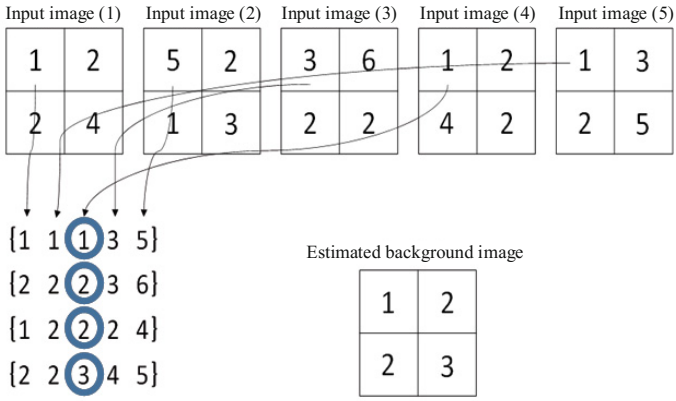


Fig. 2 Illustration of estimated background image

2.1.2 Updating Scheme for Estimated Background

There are various techniques to update the estimated background using the method of the recursive background model. Among them, the technique by using recursive median filter has little computational complexity, and is more effective. First we compare the pixel level of the estimated background which we hold with input image. If the pixel level of input image is greater than a predetermined threshold, the pixel level of estimated background image is increased by α amount. On the contrary, if the pixel level of input image is less than the threshold, the pixel level of estimated background image id decreased by α. Therefore, it update per frame that innovate estimated background image. The value of α is to be chosen by user to have an ability of update level. We describe the update scheme as shown in equation (1). The pixel level of the input image of coordinate (x, y) in time t is denoted by $I_t(x, y)$. The pixel level of estimated background image is $B_t(x, y)$. The update level is α, threshold is *T*, then it can be summarized as follows. The level of α here is often chosen relatively small fixation level in most methods such as α = 1. But we have to deal with detecting cow region as particular case when illumination suddenly changes [6].

$$B_t(x,y) = \begin{cases} B_{t-1}(x,y) + \alpha & if\ (\{I_t(x,y) - B_{t-1}(x,y)\} > T) \\ B_{t-1}(x,y) + \alpha & if\ (\{I_t(x,y) - B_{t-1}(x,y)\} < -T) \end{cases} \quad (1)$$

2.2 Foreground Extraction

Suppose that for coordinate (x, y) in time t, the luminance is $I_t(x, y)$, luminance of estimated background image is $Bt(x, y)$. If for the threshold variable T, condition type of a range detected as a foreground set is defined as

$$| I_t(x, y) - B_t(x, y) | > T \quad (2)$$

Absolute value at each coordinate of input image and estimated background image express the coordinate which is bigger than threshold. An advantage of our approach is that there is little computational complexity as can be shown experimentally. For the complexity that may occur under various environments we have to take the appropriate threshold. But if we can create the background that is correct with an estimated background image, we can expect the extracted foreground set of high precision. We show an example of the foreground set extraction in Fig 3 by using the method of a background difference [6].

Fig. 3 Example of the foreground set extraction by a background difference

2.2.1 A Noise Removal and Morphological Operation

A foreground set image is output as a binary image. For this output image, the cow region and background region are to be verified. So, domain supplement lacking in the detection process for cow region is necessary to perform noise reduction. In order to do so, we perform dilation of the morphology processing by the domain supplement that performs labeling process for the noise reduction.

3 Cow Identification

In this section we will develop an algorithm that extracts a mark domain to be used for identification from the cow region which has been extracted in section 2. Then the process to identify an individual cow will be described in the following. The general procedure is as shown in Fig 4. First, we extract region of mark that were drawn on cow. Then, we extract the shape feature. Finally, we detect similarity between query image and the image in database to identify an individual.

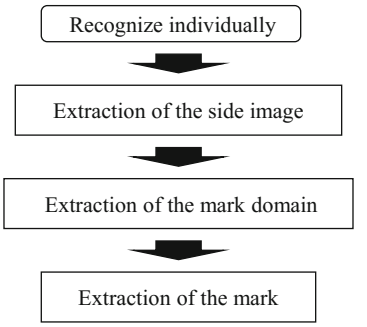


Fig. 4 Flow diagram of mark detection

3.1 ROI Detection

We detect the mark from cow region in this subsection. We show a flow diagram to mark detection. As step 1, it creates a histogram to vertical direction and horizontal directions, that we detect a region of the side of cow having many regions along the vertical axis and regions of foot are in a horizontal axis. This step prevents that the color of tag which is attached to ear of the cow is recognized as the mark. As step2, we detect region of mark by using values in an HSV color space [7]. Finally, as step3, we detect only mark only in order to delete region of the skin. In this step we calculate similarity by using feature of mark alone.

3.2 Feature Extraction

We calculate an angle from the middle point of the image, and we treat its number as feature quantity. To calculate an angle θ, we divide image into four images of fourth quadrant from first quadrant. And, we count angle of the number in range of the 24 intervals and create as a feature vector. We show below a formula to calculate the value of θ. n is shows image size.

[first quadrant] $\theta_1(^\circ) = \tan^{-1}((n/2 - j)/i) * 180/\pi$

[second quadrant] $\theta_2(^\circ) = \tan^{-1}((n/2 - i)/(n/2 - j) * 180/\pi + 90$

[third quadrant] $\theta_3(^\circ) = \tan^{-1}(j/(n/2 - i) * 180/\pi + 180$

[fourth quadrant] $\theta_4(^\circ) = \tan^{-1}(i/j) * 180/\pi + 270$

3.3 Feature Matching

For similarity, we use the Euclidean distance [8]. In this process we convert query image in the database into binary image and we compute the feature quantities. Then the Euclidean distances between feature vectors of query and images from the database are computed to make a decision of similarity.

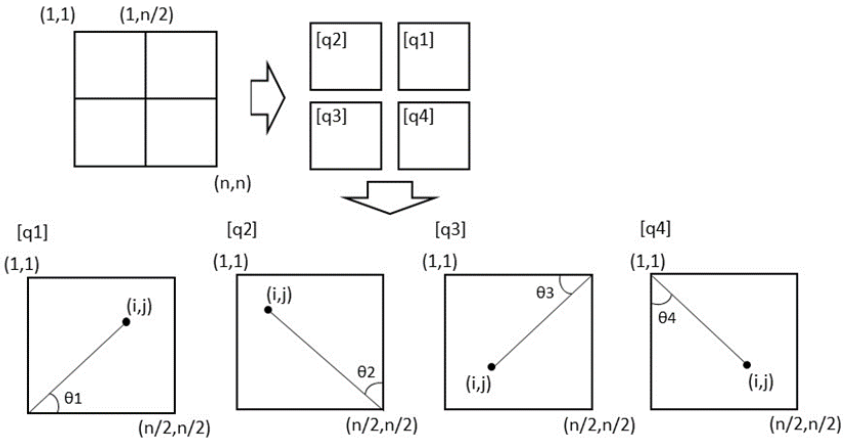


Fig. 5 Calculation of Angle θ

4 Experimental Results

For experimental setting, the images are taken in the University dairy farm. Installation location of camera is about 2 meters high from ground. Evaluation method is shown below. At first, three patterns are drawn. From the total of 30 pieces, for each 10 pieces in three pattern images are maintained in database. Then, we perform a mark detecting and the matching of the image in the database for every one mark from the image. We calculate the similar degrees between each mark image in the database and query image with the mark that we extracted from an animation image. Then we display it in a ranking depending on the degrees of similarities starting from large to small scores. We took out an image of the ranking 15^{th} place and adopted the mark that we counted most. Finally, we distinguish the cow by the combining using the mark that we adopted. In addition, we include alarm for an error identification when a combination does not exist. The experimental works are shown in Fig 6. Evaluation results are described in Table 1.

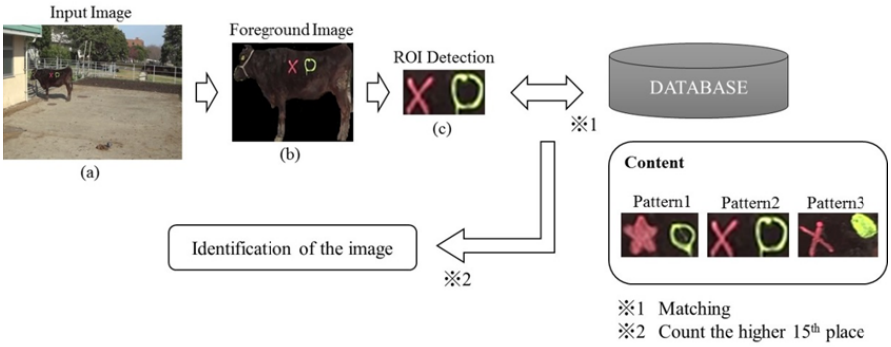


Fig. 6 Flowchart of Experimental Procedure

Table 1 Evaluation results

	Number of image				
	pattern1	pattern2	pattern3	others	Accuracy Rate(%)
pattern1	21	6	0	3	70
pattern2	0	30	0	0	100
pattern3	0	0	14	16	46
				total	78.8

5 Conclusions

We had presented a new system for detecting and identification of individual black-haired cows from video sequences. In our approach we have used only shape feature of marks. To have more accuracy, we would need to improve the proposed system by using more features including color feature. We hope this work will be done in our future works. In addition, the system which solves flexibly problems in which some parts of marks are hidden and shapes of marks are gotten out will be created.

Acknowledgment. This work is partially supported by KAKENHI 15K14844 Grant-in-Aid for Challenging Exploratory Research.

References

1. Kouichiro, H., Kobayashi, I., Kajisa, M., Kitahara, G., Fukuyama, K., Shunichi, K.: Effect of time lapse from the onset estrus detected by pedometer until insemination and synchronization for artificial insemination on calf sex ration in Japanese Black Cows, p. 94. University of Miyazaki, Department of agriculture memoir, Japan (2010) (Japanese)
2. Kawaguchi, K.: Individual Identification of Milk Cow using robust spot expression method to posture change, Mie University (Master thesis), p. 1, 2004, Japan (Japanese)
3. Chalechale, A., Naghdy, G., Premaratne, P.: Angler-Radial decomposition algorithm for sketch-based image retrieval. In: Proc. of International Conf. on Image Processing (ICIP2000), vol. 3 (2000)
4. Xie, Y.-H., Setia, L., Burkhardt, H.: Object-based Color Image Retrieval Using Concentric Circular Invariant Features. IJCSES International Journal of Computer Sciences and Engineering Systems **1**(3), July 2007
5. Suzue, N.: Edge Feature for Monochrome Image Retrieval, Kochi University of Technology, Information System Engineering (Graduate Thesis) (2010) (Japanese)
6. Mukai, Y.: The Method about Foreground Extraction Method which is Easy to Fit an Environmental Change by the Combination Long-term update and Short-term update, Osaka City University, Graduate School of Engineering (Graduate Thesis), Japan, page 13 (2011) (Japanese)
7. Pratt, W.K.: Digital Image Processing: PIK Inside, 3rd edn., vol. 14, p. 402
8. Minami, T., Nakamura, O.: Image Engineering, Corona publishing, p. 94, October 2005

Perfect Play in Miniature Othello

Yuki Takeshita, Makoto Sakamoto, Takao Ito and Satoshi Ikeda

Abstract In 1993, J. Feinstein reported that a perfect play on 6×6 board of Othello gives a 16-20 win for the second player. His result does not remain as a paper, since he reported it on the Newsletter of the British Othello Federation. In our previous paper, we found out a perfect play different from the Feinstein. They have the same score, but it will not be a proof of the Feinstein is a perfect play. In this paper, we confirm that the sequence reported by Feinstein is one of the perfect plays. In addition, we introduce one of perfect plays in each boards of 4×4, 4×6, 4×8 and 4×10. From these results, we discuss the feature of the Othello larger than or equal to 8×8 board.

Keywords Combinatorial theory · Combinatorial optimization · Perfect play · Rectangular Othello

1 Introduction

The Othello is a board game derived from Reversi, and it is devised by Goro Hasegawa (JPN) in 1973. Othello rules are completely unified, whereas the

Y. Takeshita(✉)
Graduate School of Engineering, Miyazaki University, 1-1,
Gakuen Kibanadai Nishi, Miyazaki 889-2192, Japan
e-mail: hf11031@student.miyazaki-u.ac.jp

M. Sakamoto · S. Ikeda
Department of Computer Science and System Engineering,
Miyazaki University, 1-1, Gakuen Kibanadai Nishi, Miyazaki 889-2192, Japan
e-mail: {sakamoto,bisu}@cs.miyazaki-u.ac.jp

T. Ito
Graduate School of Engineering, Hiroshima University, 1-4-1, Kagamiyama,
Higashihiroshima 739-8527, Japan
e-mail: itotakao@hiroshima-u.ac.jp

T.T. Zin et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 388,
DOI: 10.1007/978-3-319-23207-2_28

Reversi has a lot of local rules. World championships have been held every year since 1977, tournament is held in Japan, the United States and European countries.

This game is categorized into two-player zero-sum finite deterministic games of perfect information [1]. Games in this class are possible to look ahead in theory, thus if both players choose constantly the best move, these are classified into a win, loss or draw game for the first player [2]. Standard 8×8 board of Othello has not been solved because of too large positions; according to some strong programs, it seems the draw theory is strong [3]. In 1993, Joel Feinstein reported that a perfect play on 6×6 board of Othello gives a 16-20 win for the second player [4](Fig. 1). In our previous paper [5], we found out a perfect play different from the Feinstein. They have the same score, but it will not be a proof of the Feinstein is a perfect play.

In this paper, we confirm that the sequence reported by Feinstein is one of the perfect plays. In addition, we introduce one of perfect plays in each boards of 4×4, 4×6, 4×8 and 4×10. From these results, we discuss the feature of the Othello larger than or equal to 8×8 board.

28	9	8	7	31	32
22	14	1	4	17	16
27	13	○	●	6	15
19	2	●	○	5	12
24	23	3	10	26	29
21	18	20	11	25	30

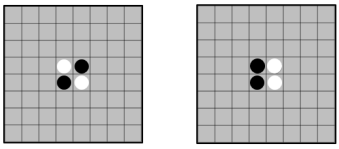
Fig. 1 Feinstein's perfect play.

2 Othello[1]

First of all, we will introduce the rules of Othello. See Fig. 2 (a). The game always begins with this setup. In case of the Reversi, the first two moves by each player are in the four central squares of the board, thus there may be a parallel as Fig. 2 (b). One player uses the black side of the pieces (circular chips), the other the white sides. Black always moves first.

Both players put the pieces of own color to an empty board in turn. A player's move consists of outflanking his opponent's the pieces. Then, he flip outflanked the pieces to his color. To outflank means to place the piece on the board so that his opponent's rows of the piece are bordered at each end by the piece of his color. If a player cannot make a move that flips at least one of his opponent the pieces, then he has to pass. If he is able to make a valid move however, then passing is not allowed. The game ends when neither player can make a valid move. The winner is the player who has more the pieces than his opponent.

[1] Othello is a registered trademark.



(a) Othello board. (b) Reversi board.

Fig. 2 Starting position.

3 Computer Othello

The making of the thinking routines is indispensable in studying perfect analysis of the board game. This is because the end-game routine is the perfect analysis exactly, and an evaluation function in the middle-game routine is available for the ordering of the search in perfect analysis.

In addition, end-game is classified into solver for WLD (win/loss/draw) score and solver for exact score. Both are the perfect analysis, but there is a difference in the evaluation of the end. In solver for exact score, the end is evaluated with the piece difference. However, it is necessary to consider if one was wiped out in the middle; for example, in 4×8 board, the evaluation value at **26-0** must be **+32**. This routine can find best one move. In solver for WLD score, the end is evaluated in three ways win, loss, and draw. It generates relatively many pruning (Alpha-Beta Pruning [6]) if the range of the evaluation value is small. Therefore, an execution time for the exact score is estimated to be several times of the execution time for the WLD score.

3.1 Speed Up of the Program

Currently, our program read approximately 1.5-2 million moves per second. There are some ideas for this. At first, we implemented doubly-linked list that stores the empty pieces. This function reduce search cost so as to go to the end-game. From this, the search speed has approximately doubled. Next, our program had a function to count number of the pieces in specified color, but we have removed it. Instead we adjusted at flips and unflips by adding a variable that stores number of the pieces into the structure. This effect was approximately 1.5 times.

4 Perfect Play

A perfect play in computer Othello is a sequence when both players choose constantly the best move. In this paper, we confirm the result of Feinstein in 6×6 Othello board by reading his perfect play at the very first in the perfect analysis solver.

5 Experiments

In this section, we confirm that the sequence reported by Feinstein is one of the perfect plays. In addition, we introduce one of perfect plays in each Othello or Reversi boards of 4×4, 4×6, 4×8 and 4×10.

5.1 6×6 Othello Board

In 6×6 Othello board, the second player win. Final result of the perfect play is ***Black: 16, White: 20***. It is shown to Fig. 3. It was passed at move 31. In view of Fig. 3 and Fig. 1, it is seen that our perfect play fully coincident with Feinstein's.

28	9	8	7	32	33
22	14	1	4	17	16
27	13	W	B	6	15
19	2	B	W	5	12
24	23	3	10	26	29
21	18	20	11	25	30

Fig. 3 Sequence of the perfect play (6×6 Othello board).

5.2 4×4 Othello Board

In 4×4 Othello board, the second player win. Final result of the perfect play is ***Black: 3, White: 11***. It is shown to Fig. 4. The elements mean that **B** is the black pieces and **W** is the white pieces. Also, it was no pass in this game.

2	3	4	7
1	W	B	
6	B	W	
9	8	10	5

Fig. 4 Sequence of the perfect play (4×4 Othello board).

5.3 4×4 Reversi Board

In 4×4 Reversi board, the second player win. Final result of the perfect play is ***Black: 6, White: 9***. It is shown to Fig. 5. It was passed in move 10.

2	6		1
3	B	W	9
4	B	W	8
7	11	12	5

Fig. 5 Sequence of the perfect play (4×4 Reversi board).

5.4 4×6 Othello Board

In 4×6 Othello board, the first player win. Final result of the perfect play is ***Black: 20, White: 4***. It is shown to Fig. 6. It was passed at move 18.

3	2	1	10	6	11
16	5	W	B	8	19
17	14	B	W	9	20
15	21	4	13	7	12

Fig. 6 Sequence of the perfect play (4×6 Othello board).

5.5 4×6 Reversi Board

In 4×6 Reversi board, the first player win. Final result of the perfect play is ***Black: 21, White: 3***. It is shown to Fig. 7. It was passed at move 20.

15	16	5	4	1	17
19	6	B	W	12	21
14	2	B	W	8	10
13	11	3	7	18	9

Fig. 7 Sequence of the perfect play (4×6 Reversi board).

5.6 4×8 Othello Board

In 4×8 Othello board, the first player win. Final result of the perfect play is ***Black: 28 White: 0***. It is shown to Fig. 8. It was passed at move 12, 24 and 26.

15	3	2	1	4	5	19	
20	14	10	W	B	8	22	27
25	17	13	B	W			
21	16	9	6	11	7	18	23

Fig. 8 Sequence of the perfect play (4×8 Othello board).

5.7 4×8 Reversi Board

In 4×8 Reversi board, the first player win. Final result of the perfect play is ***Black: 28 White: 0***. It is shown to Fig. 9. It was passed at move 20, 22 and 26.

	5	2	3	4	1	10	11
25	24	12	B	W	9	16	18
		23	B	W	15	17	27
	21	6	13	8	7	14	19

Fig. 9 Sequence of the perfect play (4×8 Reversi board).

5.8 4×10 Othello Board

In 4×10 Othello board, the first player win. Final result of the perfect play is ***Black: 39, White: 0***. It is shown to Fig. 10. It was passed at move 19, 28, 34, 36, 38 and 40.

21	20	3	2	1	4	5	13	30	31
24	17	16	14	W	B	10	22	29	32
25	26	23	15	B	W	18	39	35	33
27	37	8	7	6	11	9	12	41	

Fig. 10 Sequence of the perfect play (4×10 Othello board).

5.9 4×10 Reversi Board

In 4×10 Reversi board, the first player win. Final result of the perfect play is ***Black: 32, White: 0***. It is shown to Fig. 11. It was passed at move 24, 26 and 28.

		5	2	3	4	1	20	21	
	29	15	23	B	W	25	17		27
31	30	16	14	B	W	8	10	18	22
		13	6	11	7	12	9		19

Fig. 11 Sequence of the perfect play (4×10 Reversi board).

5.10 6×6 Reversi Board

In 6×6 Reversi board, the second player win. Final result of the perfect play is ***Black: 17, White: 19***. It is shown to Fig. 12. It was no pass.

31	9	8	5	15	25
27	16	12	2	24	26
14	13	B	W	1	4
19	11	B	W	3	17
29	28	10	6	18	23
30	32	20	7	21	22

Fig. 12 Sequence of the perfect play (6×6 Reversi board).

5.11 Execution Data

Table 1 shows an execution data of the perfect analysis in each Othello board. "Position" means number of the final position, "Time" means execution time and "Result" means final results of the perfect play.

Table 1 Execution results in each Othello board.

	Position	Time	Result
4×4	218	0.001s	LOSS (-8) B: 3, W: 11
4×6	139,803	0.1s	WIN (+16) B: 20, W: 4
4×8	294,430,331	2m15s	WIN (+32) B: 28, W: 0
4×10	1,195,804,922,641	6d6h22m	WIN (+40) B: 39, W: 0
6×6	884,392,099,420	5d12h16m	LOSS (-4) B: 16, W: 20

Table 2 shows the data in each Reversi board. Seeing the "Result" axis, it can obtain the similar results as the Othello board.

Table 2 Execution results in each Reversi board.

	Position	Time	Result
4×4	524	0.001s	LOSS (-3) B: 6, W: 9
4×6	274,549	0.15s	WIN (+18) B: 21, W: 3
4×8	299,987,758	2m12s	WIN (+32) B: 28, W: 0
4×10	842,204,125,277	4d12h22m	WIN (+40) B: 32, W: 0
6×6	1,628,664,185,199	8d12h42m	LOSS (-2) B: 17, W: 19

6 Consideration

See Fig. 13 and Fig. 14. These show acquisition rate of first move in the perfect play. The horizontal axis represents each board and the vertical one represents acquisition rate of first move relative to the total pieces of final results in the perfect play. Therefore, above the 50% shows the first player win. Below shows the second player win. Additionally, the elements in Fig. 13 and Fig. 14 mean that the circles are Othello board and the triangles are Reversi board.

In Fig. 13, it is shown that the acquisition rate is increased in the transition from the 4×4 board to the 6×6 board. In Fig. 14, even in rectangular Othello, the acquisition rate is similarly increased. Based on these considerations, we guess the Othello have feature that gives the first player an advantage according to extended board size. Also, considering the increased range in Fig. 13, the acquisition rate is a high possibility if not the first player win. Moreover, there is a high possibility that first move wins in boards larger than or equal to 10×10.

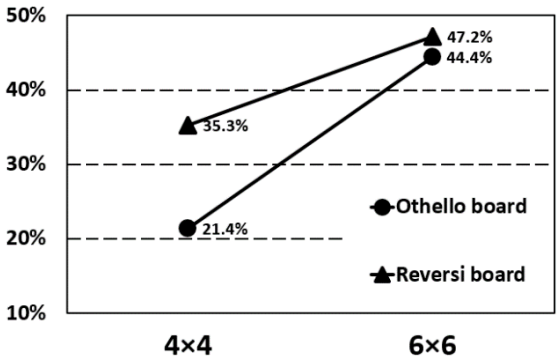


Fig. 13 Acquisition rate of first move (Square).

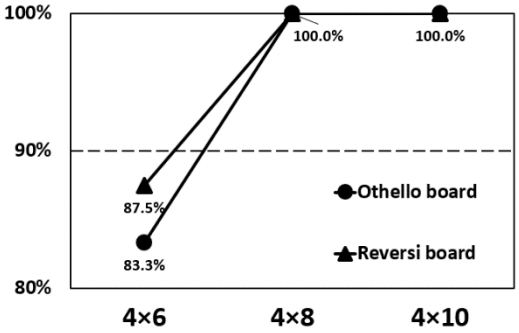


Fig. 14 Acquisition rate of first move (Rectangle).

Besides, Fig. 15 shows the transition in number of the final positions in Othello board. The horizontal axis shows each board size and the vertical one represents number of the final positions by the exponent. The elements in Fig. 15 mean that the circles are square boards and the triangles are rectangular boards.

First, in rectangle board, it can be guessed that number of the final positions in 4×12 board is about 10^4 times in 4×10 board. Given that the execution time in 4×10 board is approximately six days, the execution time in 4×12 board can be guessed about sixty thousand.

Secondly, in square board, it has increased about 10^{10} times in the transition from the 4×4 board to the 6×6 board. Therefore, number of the final positions in 8×8 board can be guessed to be about 10^{22} at least. Given that the execution time in 6×6 board is approximately five days and a half, we know that perfect analysis in 8×8 board is impossible now.

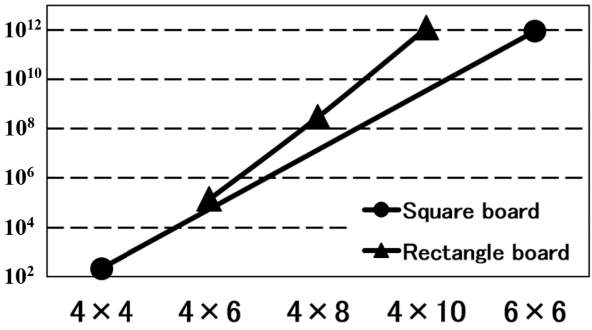


Fig. 15 Number of the final positions in Othello board.

7 Future Work

In this paper, we confirmed that the sequence reported by Feinstein is one of the perfect plays. We will try the perfect analysis in more extended boards (4×12,

4×14 and 6×8) to confirm that our consideration is correct. In order to solve them, we plan the use of supercomputers after improvements of our solver.

7.1 Improvement

Move ordering performs well for Alpha-Beta pruning. Now, because our program uses low quality it, it is necessary to improve. We have to implement a hash table to cut the boards with duplicate and symmetry. However, it will slow down the search because it takes time for access to it.

Acknowledgements. The authors thank anonymous referees for their useful comments. This work was supported by JSPS KAKENHI Grant Number 24510217.

References

1. Neumann, J.V., Morgenstern, O.: Theory of Games and Economic Behavior. Princeton University Press (1944)
2. Schaeffer, J., Burch, N., Björnsson, Y., Kishimoto, A., Müller, M., Lake, R., Lu, P., Sutphen, S.: Checkers Is Solved. Science **317,** 1518–1522 (2007)
3. Rose, B.: Othello - A Minute to Learn… A Lifetime to Master (2005). www.ffothello.org/livres/othello-book-Brian-Rose.pdf (Retrieved June 18, 2015)
4. British Othello Federation: Forty Billion Nodes Under The Tree - The Newsletter of the British Othello Federation, pp. 6–8 (1993). www.britishothello.org.uk/fbnall.pdf (Retrieved June 18, 2015)
5. Takeshita, Y., Ikeda, S., Sakamoto, M., Ito, T.: Perfect analysis in miniature othello. In: Proceedings of the 2015 International Conference on Artificial Life and Robotics, p. 39 (2015)
6. Knuth, D.E., Moore, R.W.: An Analysis of Alpha-Beta Pruning. Artificial Intelligence **6**, 293–326 (1975)

Part IV
Circuits and Signal Processing with Engineering Application

The Development of the Nano-Mist Sprayer and Its Application to Agriculture

Shugo Kaminota, Koichi Tanno, Hiroki Tamura and Kiyoto Kawasaki

Abstract In this paper, we describe the development of the nano-mist sprayer using the Venturi effect. This sprayer consists of AC or DC pump for Venturi effect, cylinder for causing Venturi effect, control circuit and separation plate. The particle diameter of this nano-mist sprayer is a few hundred nanometers or less. Furthermore, the particle size can be controlled by air pressure of the pump and shape of the separation plate. This sprayer can be applied to agriculture, medical, health care, etc. In this paper, we show an evaluation result of the mist particle diameter, firstly. Next, in order to confirm the usefulness of the sprayer, we apply to the agricultural field; the repellent spray was done by the fabricated sprayer. The experimental results are reported in this paper.

Keywords Nano-mist sprayer · Venturi effect · Particle diameter · Repellent spray

S. Kaminota
Department of Electrical and Electronic Engineering, Graduate School of Engineering, University of Miyazaki, 1-1, Gakuenkibanadai-nishi, Miyazaki 889-2192, Japan

K. Tanno(✉)
Department of Electrical and Systems Engineering, Faculty of Engineering, University of Miyazaki, 1-1, Gakuenkibanadai-nishi, Miyazaki 889-2192, Japan
e-mail: tanno@cc.miyazaki-u.ac.jp

H. Tamura
Department of Environmental Robotics, Faculty of Engineering, University of Miyazaki, 1-1, Gakuenkibanadai-nishi, Miyazaki 889-2192, Japan

K. Kawasaki
Wisdom Co.Ltd., 569-3, Minami-takanabe, Takanabe-cho, Koyu-gun, Miyazaki 884-0003, Japan

T.T. Zin et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 388,
DOI: 10.1007/978-3-319-23207-2_29

1 Introduction

In recent years, traceability of agricultural crops is important for safety of foods. Many farmers try to reduce the total amount of repellent and record the consumption of the repellent to keep the safety of foods [1]-[4]. Furthermore, the population of the primary industry in Japan decreases and the aging population of farmers are rapidly increases. Therefore, it is strongly required to develop the new devices to automatically treat and manage the repellents with reducing the consumption of them.

In this paper, we propose the new device based on Venturi effect to spray the repellents. In the past, there are some mist sprayers with similar mechanism, however, the particle diameter of the sprayed mist is relatively large and vary widely, so that it is difficult to diffuse the repellents extensively and the mist of the large diameter adhere around the sprayer as droplets. On the other hand, the proposed device has an advantage that the particle diameter of the sprayed mist is several hundred nanometers or less. This particle diameter is much smaller than that of conventional sprayers. Therefore, it is possible to make diffuse the sprayed mist extensively in a short time.

Firstly, the mechanism of the proposed sprayer is explained, next, the diameter of the sprayed mist of the proposed sprayer is evaluated. Lastly, we apply the proposed device to spray repellents in vinyl greenhouse, and the experimental results are shown in this paper.

2 Proposed Sprayer (Nano-Mist Sprayer)

The proposed sprayer, which we call nano-mist sprayer, is shown in Figure 1. The simplified component blocks of the proposed sprayer are shown in Figure 2. This sprayer consists of three blocks; electronic control block, main body block and cylinder block. The proposed sprayer is based on the Venturi effect, which is built in the cylinder block.

In the electronic control block, PIC (Peripheral Interface Controller) which implement timer function and the control of AC pump are included. The main body block includes AC pump to feed the compressed air to cylinder block. Figure 3 shows the cross-section view of the cylinder block. The cylinder block consists of inner tube, outer tube and separation plate. The inner tube connects to the AC pump at the bottom of the inner tube. The bottom of the outer tube is directly connected to the storage tank for the sprayed liquid.

Next, we describe the mechanism for generating the mists of the nano-order particle diameter. When the compressed air get pumped to the inner tube from AC pump, the negative pressure is occurred between inner and outer tubes, that is to say, Venturi effect is occurred. Therefore, the liquid, which is supplied from storage tank, is absorbed toward the top of the tube. The absorbed liquid is upward sprayed by the compressed air. The sprayed mists collide against the separation plate, and the mists of the large particle diameter are dropped to the storage tank. Thus, it contributes to

reduce the consumption of the sprayed liquid. On the other hand, the fine mists are sprayed from spray nozzle shown in Figure 3. As the results, the mist size is a few hundred nanometers or less. In this way, the separation plate performs as a filter.

Fig. 1 The nano-mist sprayer

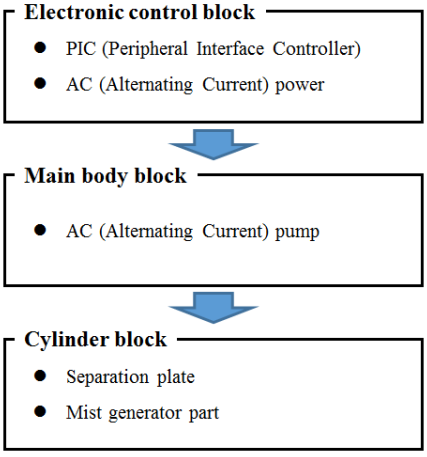


Fig. 2 The component blocks

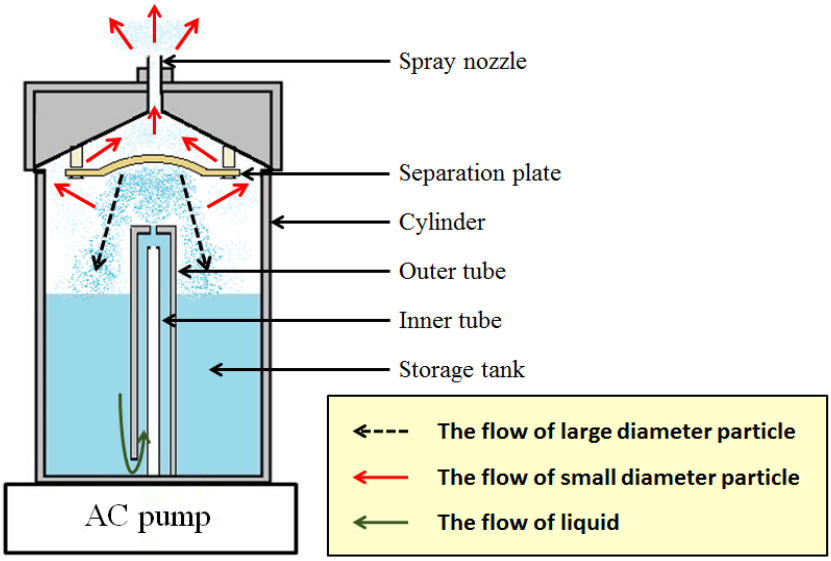


Fig. 3 The cross-section view of the cylinder block

3 Evaluation of Particle Diameter

In this chapter, we describe the evaluation of the particle diameter of the sprayed mist of the proposed sprayer. In this evaluation, "Spraytec" produced by Malvern Instruments Ltd is used to measure particle diameter. In some cases, the sprayed mists are bound each other, therefore, we evaluated the particle diameter of the sprayed mists in two conditions; the distances from the spray nozzle (L to the measuring point) are 25 mm and 100 mm (See Figure 4). This evaluation is based on the number of particles.

Figure 5 shows the evaluation results of the particle diameter and the cumulative ratio. The bar and line charts mean the ratio of each particle diameter and the cumulative ratio, respectively. In this evaluation results, we could confirm that the particle diameter of the sprayed mists is less than 1 μm. The particle diameter has a tendency to become larger as L increases. The peak values of the particle diameter under the condition of $L = 25$ mm and $L = 100$ mm are 185 nm and 215 nm, respectively. When the cumulative ratio is equal to 90 %, the particle diameter is less than 340 nm under the both conditions. From this evaluation, we could confirm the separation function of the separate plate of the proposed sprayer.

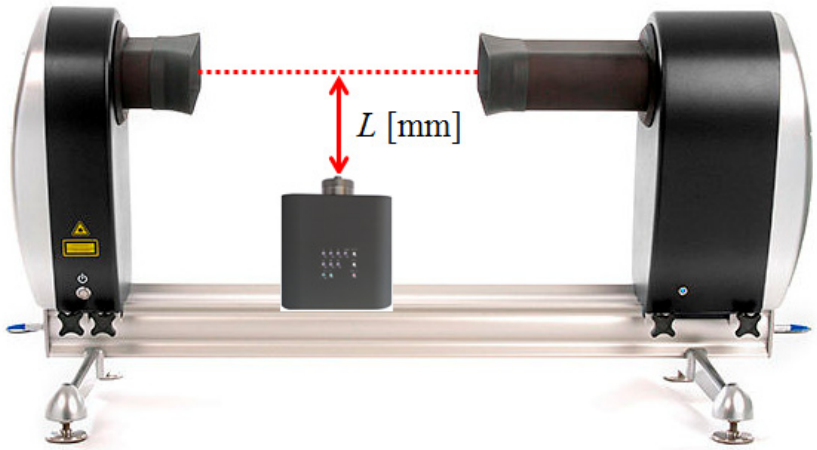


Fig. 4 Picture of Spraytec

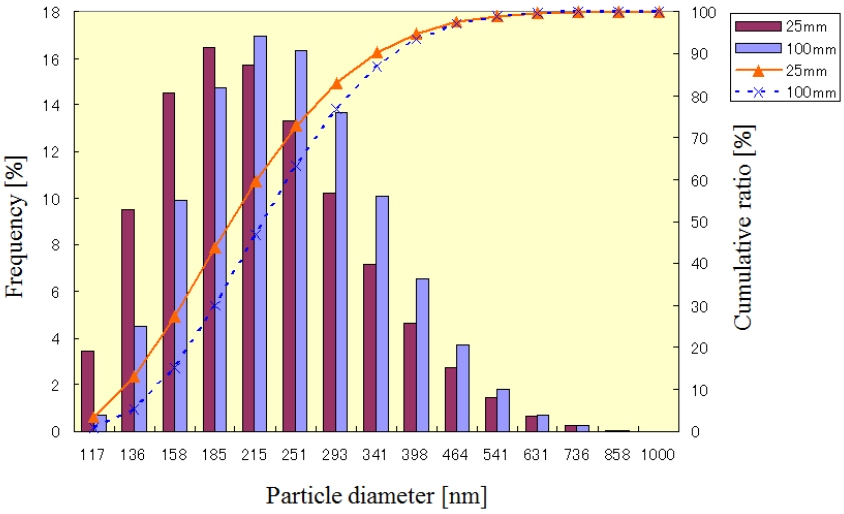


Fig. 5 The evaluation results of the particle diameter and the cumulative ratio

4 Application to Agriculture

In this chapter, we report the experiment results of the repellent spray to agricultural crops as the application of the proposed sprayer. This experiment was done in the vinyl greenhouses for green peppers. For comparison, we prepared the two vinyl greenhouses; the repellent was not sprayed in vinyl greenhouse A but sprayed in vinyl greenhouse B. The areas of the vinyl greenhouses A and B are 1000 m^2 and 250 m^2, respectively and the distance between them is 5 m. The repellent was sprayed in the vinyl greenhouse B for 10 days (100cc by one day).

Fig. 6 Agricultural crops in the vinyl greenhouse A

Fig. 7 Agricultural crops in the vinyl greenhouse B

Figures 6 and 7 show photos of the green peppers in vinyl greenhouse A which suffer from insect damage, and B which is protected from insect damage, respectively. From these results, we could confirm that the insect damage is quite different between vinyl greenhouses A and B. Furthermore, we sprayed the repellent in the vinyl greenhouse A after the foregoing experiment. As the result, we confirmed that it was possible to suppress the insect damage to agricultural crops like the vinyl greenhouse B. From these experimental results, it is very useful to apply the proposed sprayer for the repellent spray.

5 Conclusion

In this paper, we have proposed the nano-mist sprayer. The proposed sprayer has advantage that the particle diameter is less than a few hundred nanometers or less. Next, we have applied the proposed sprayer to repellent spray in the vinyl greenhouse. From the experiment results, we could confirm the effectiveness of the repellent spray using the proposed sprayer.

The mist diameter control of the proposed sprayer and the dependence of the particle diameter with sprayed liquid are the future work.

Acknowledgments This work was supported by JSPS KAKENHI Grant Number 24658213.

References

1. Auernhammer, H.: The Role of Mechatronics in Crop Product Traceability. Agricultural Engineering International: the CIGR Journal of Scientific Research and Development **IV**, October 2002
2. Miller, P.C.: Patch Spraying: Future Role of Electronics in Limiting Pesticide Use. Pest Manag. Sci. **59**(5), 566–574 (2003)
3. De Rudnicki, V., Ruelle, B., Douchin, M., Maurel, V.B.: Embedded ICT technology on sprayers in order to reduce water pollution; the aware project. In: 8th Fruit Nut and Vegetable Production Engineering Symposium, FRUTIC 2009, Concepcion, Chile January 2009, p. 8 (2009)
4. Sven, P.: Specification, Design and Evaluation of an Automated Agrochemical Traceability System. Ph.D. Thesis, Cranfield University, U.K. (2009)

Low Offset Voltage Instrumentation Amplifier by Using Double Chopper Stabilization Technique

Makoto Sada, Koichi Tanno, Masaya Shimoyama, Zainul Abidin, Hiroki Tamura and Takako Toyama

Abstract In this paper, we propose a low offset voltage instrumentation amplifier (INA) for biological signal. The proposed circuit consists of the INA and double chopper stabilization circuit for achiving low offset voltage. The proposed circuit was evaluated through HSPICE using 1P 2M 0.6μm CMOS process. The offset voltage could be reduced to 4.1[mV] with the power consumption of 75.9[μW].

Keywords Instrumentation amplifier · Double chopper stabilization technique · Offset voltage

1 Introduction

In recent years, the interface devices between human and equipment have been actively developed. For example, a lot of application using biological signal to health care and electronic machines systems are reported [1], [2]. This kind of systems need to amplify the signals with removing the noises because the biological signals are

M. Sada · M. Shimoyama
Department of Electrical and Electronic Engineering, Graduate School of Engineering, University of Miyazaki, 1-1, Gakuenkibanadai-nishi, Miyazaki 889-2192, Japan

K. Tanno(✉) · T. Toyama
Department of Electrical and Systems Engineering, Faculty of Engineering, University of Miyazaki, 1-1, Gakuenkibanadai-nishi, Miyazaki 889-2192, Japan
e-mail: tanno@cc.miyazaki-u.ac.jp

Z. Abidin
Department of Materials and Informatics, Interdisciplinary Graduate School of Agriculture and Engineering, University of Miyazaki, 1-1, Gakuenkibanadai-nishi, Miyazaki 889-2192, Japan

H. Tamura
Department of Environmental Robotics, Faculty of Engineering, University of Miyazaki, 1-1, Gakuenkibanadai-nishi, Miyazaki 889-2192, Japan

T.T. Zin et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 388,
DOI: 10.1007/978-3-319-23207-2_30

very small amplitude and low frequency band signals. Therefore, instrumentation amplifier (INA) is utilized as preamplifier for biological signal because the INA has the advantages of high-input-impedance, high-gain and high-CMRR. However, the circuit has disadvantage of large offset voltage because the flicker noise occurs in the MOSFET. This flicker noise is inversely proportional to frequency, so that we cannot discriminate the flicker noise and biological signals if the circuit is implemented by using MOSFETs. Therefore, it needs to remove the offset voltage and the flicker noise. It is well-known that chopper stabilization technique is very useful for removing the offset voltage and the flicker noise. In the reference [3], chopper stabilization technique was applied to INA. The offset voltages of the non-inverting amplifier block (first block) could be removed, however, it remained the problem that the offset voltage of the differential block (second block) was not removed.

In this paper, we propose the low offset voltage INA for biological signals. The proposed circuit consists of the INA and double chopper stabilization circuit for removing the offset voltage and the flicker noise of both blocks. We designed and evaluated the proposed INA using 1P 2M 0.6μm CMOS process through HSPICE. The simulation results are reported in this paper.

This paper is organized as follows. In Section 2, chopper stabilization technique and the problems of the conventional circuit are discussed. In Section 3, the proposed circuit and double chopper stabilization technique are shown. Next, the simulation results of the proposed circuit are presented in Section 4. Finally, Section 5 concludes this paper.

2 Conventional INA Used Chopper Stabilization Technique

Figure 1 shows the conventional circuit used chopper stabilization technique in INA. $V_{os1,2,3,4}$ are each input-reffered offset voltages. Figure 2 shows chopper switches (SW) for chopper stabilization technique. The switch is composed by MOSFETs as shown in Figure 3.

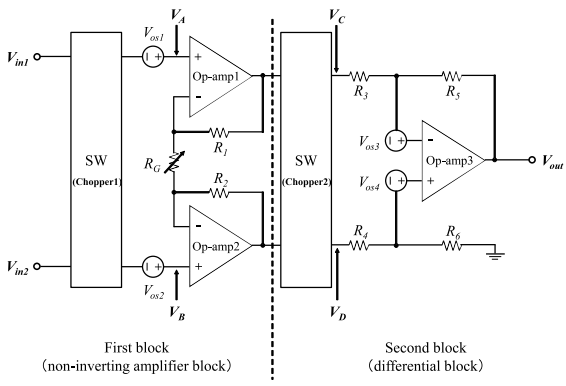


Fig. 1 The conventional circuit

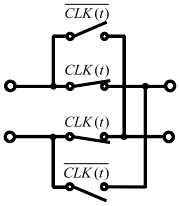


Fig. 2 Chopper switches (SW)

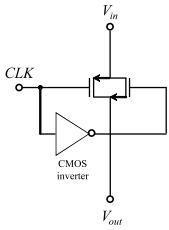


Fig. 3 CMOS switch

Firstly, we defined the input signals as $V_{in1} = V_{in} \sin(\omega_{in} t)$, $V_{in2} = -V_{in} \sin(\omega_{in} t)$. Furthermore, CLK, $\overline{CLK}$ (see Fig. 2) use clock signals. Each clock signals are expanded using the Fourier series, therefore, we can get the following equations.

$$CLK = \frac{1}{2} + g_1(t) \tag{1}$$

$$\overline{CLK} = \frac{1}{2} - g_1(t) \tag{2}$$

Where g_1(t) is defined as follows.

$$g_1(t) = \frac{2}{\pi} \sum_{n=1}^{\infty} \frac{1}{2n-1} \sin\{(2n-1)\omega_c t\} \tag{3}$$

V_{in1} and V_{in2} are modulated by SW (Chopper1 shown in Fig.1). This modulated signals are obtained by the sum of $V_{in1} \cdot CLK$ and $V_{in2} \cdot \overline{CLK}$. Finally, V_A and V_B are the sum of these modulated signals and the offset voltages. Therefore, V_A and V_B can be given by

$$V_A = 2g_1(t) V_{in} \sin(\omega_{in} t) + V_{os1} \tag{4}$$

$$V_B = -2g_1(t) V_{in} \sin(\omega_{in} t) + V_{os2} \tag{5}$$

V_A and V_B are amplified by Op-amp1 and Op-amp2 which have the gain of $1+2R_1/R_G$. After that, each amplified signals pass through the SW (Chopper2). At this moment, the input signals V_{in1} and V_{in2} are demodulated by SW (Chopper2), however, the offset voltages of Op-amp1 and Op-amp2 are modulated (not demodulated). Therefore, the offset voltages are moved to high frequency band from DC. Based on these discussion, V_C and V_D can be given by

$$V_C = \left(1 + 2\frac{R_1}{R_G}\right)\{V_{in} \sin(\omega_{in} t) + g_1(t) V_{os1} - g_1(t) V_{os2}\} + \frac{V_{os1} + V_{os2}}{2} \tag{6}$$

$$V_D = -\left(1 + 2\frac{R_1}{R_G}\right)\{V_{in} \sin(\omega_{in} t) + g_1(t) V_{os1} - g_1(t) V_{os2}\} + \frac{V_{os1} + V_{os2}}{2} \tag{7}$$

V_C and V_D are amplified and subtracted by Op-amp3 in the second block. Therefore, V_{out} is derived as follows.

$$V_{out} = -2\frac{R_5}{R_3}\left(1 + 2\frac{R_1}{R_G}\right)\{V_{in}\sin(\omega_{in}t) + g_1(t)V_{os1} - g_1(t)V_{os2}\} - \frac{R_5}{R_3}(V_{os3} - V_{os4}) \quad (8)$$

From Eq. 8, we can separate the input signals and the offset voltages of Op-amp1 and Op-amp2 by using low-pass filter. However, it can be seen that the offset voltages of Op-amp3 (the second term in Eq. 8) remains and are not modulated. Therefore, we cannot separate the input signals and the offset voltages of Op-amp3. In many cases, in the second block, the gain of 0[dB] is widely used in order to reduce the influence of the offset voltages. However, the gain of INA is drastically limited in this case.

3 Proposed INA with Double Chopper Stabilization Technique

Figure 4 shows the proposed INA. In the proposed INA, double chopper stabilization technique is utilized. This circuit consists of INA and two pairs of chopper switches (SW1 and SW2).

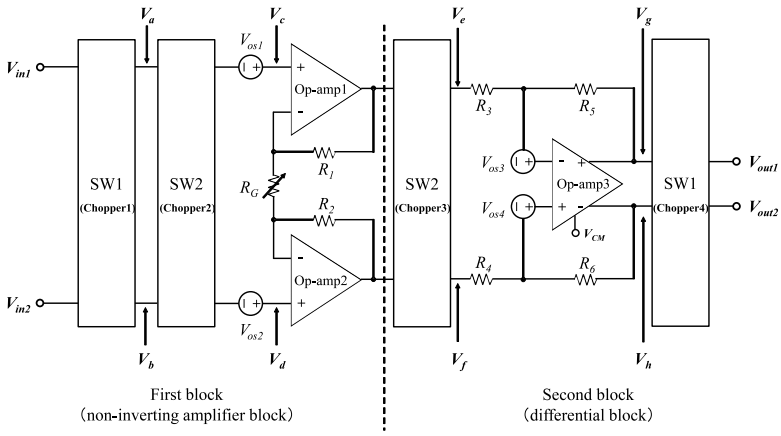


Fig. 4 The proposed INA utilized double chopper stabilization technique

In this section, we discuss about double chopper stabilization technique.

In the same manner as Sec.2, the clock signals are defined as follows.

$$CLK1 = \frac{1}{2} + g_2(t) \tag{9}$$

$$\overline{CLK1} = \frac{1}{2} - g_2(t) \tag{10}$$

$$CLK2 = \frac{1}{2} + g_3(t) \tag{11}$$

$$\overline{CLK2} = \frac{1}{2} - g_3(t) \tag{12}$$

where $CLK1$ and $CLK2$ are used for SW1 and SW2, respectively and, $g_2(t)$ and $g_3(t)$ are as follows.

$$g_2(t) = \frac{2}{\pi} \sum_{n=1}^{\infty} \frac{1}{2n-1} \sin\{(2n-1)\omega_{c1}t\} \tag{13}$$

$$g_3(t) = \frac{2}{\pi} \sum_{n=1}^{\infty} \frac{1}{2n-1} \sin\{(2n-1)\omega_{c2}t\} \tag{14}$$

Since V_{in1} and V_{in2} are modulated by SW1 (Chopper1), V_a and V_b can be given by

$$V_a = 2g_2(t)V_{in} \sin(\omega_{in}t) \tag{15}$$

$$V_b = -2g_2(t)V_{in} \sin(\omega_{in}t) \tag{16}$$

And then, V_a and V_b are also modulated by SW2 (Chopper2), and the offset voltages are added, therefore, V_c and V_d are as follows.

$$V_c = 4g_2(t)g_3(t)V_{in} \sin(\omega_{in}t) + V_{os1} \tag{17}$$

$$V_d = -4g_2(t)g_3(t)V_{in} \sin(\omega_{in}t) + V_{os2} \tag{18}$$

In this way, the input signals are modulated twice as illustrated in Figure 5.

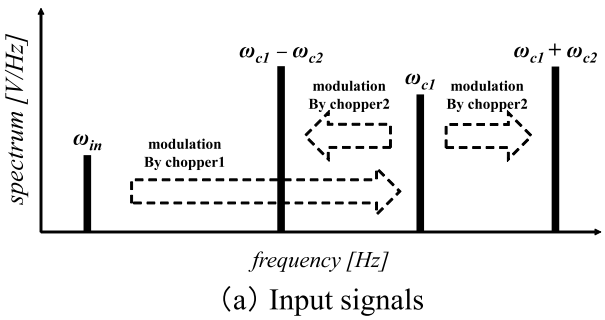


Fig. 5 The flow of the input signals in the first block

V_c and V_d are amplified by Op-amp1 and Op-amp2 which have the gain of $1+2R_1/R_G$. After that, each amplified signals pass through the SW2 (Chopper3). Because the input signals are modulated twice by Chopper1 and Copper2, and the modulated signals are demodulated by Chopper3, the frequency of the signals V_e and V_f are equal to V_a and V_b (but amplitudes are different). On the other hand, offset voltages V_{os1} and V_{os2} are modulated by Chopper3.
As the results, V_e and V_f can be given by

$$V_e = \left(1 + 2\frac{R_1}{R_G}\right)\{2g_2(t)V_{in}\sin(\omega_{in}t) + g_3(t)V_{os1} - g_3(t)V_{os2}\} + \frac{V_{os1} + V_{os2}}{2} \quad (19)$$

$$V_f = -\left(1 + 2\frac{R_1}{R_G}\right)\{2g_2(t)V_{in}\sin(\omega_{in}t) + g_3(t)V_{os1} - g_2(t)V_{os2}\} + \frac{V_{os1} + V_{os2}}{2} \quad (20)$$

Since V_e and V_f are subtracted and amplified by the second block, V_g and V_h are derived as follows.

$$V_g = -\frac{R_5}{2R_3}\left(1 + 2\frac{R_1}{R_G}\right)\{4g_2(t)V_{in}\sin(\omega_{in}t) + 2g_3(t)V_{os1} - 2g_3(t)V_{os2}\} - \frac{R_5}{2R_3}(V_{os3} - V_{os4}) \quad (21)$$

$$V_h = \frac{R_5}{2R_3}\left(1 + 2\frac{R_1}{R_G}\right)\{4g_2(t)V_{in}\sin(\omega_{in}t) + 2g_3(t)V_{os1} - 2g_3(t)V_{os2}\} + \frac{R_5}{2R_3}(V_{os3} - V_{os4}) \quad (22)$$

V_g and V_h pass through the SW1(Chopper4). The input signals are further demodulated by Chopper4. In the meantime, all of the offset voltages are further modulated by Chopper4. V_{out1} and V_{out2} are derived as follows.

$$V_{out1} = -\frac{R_5}{R_3}\left(1 + 2\frac{R_1}{R_G}\right)\{V_{in}\sin(\omega_{in}t) + 2g_2(t)g_3(t)V_{os1} - 2g_2(t)g_3(t)V_{os2}\} - \frac{R_5}{R_3}g_2(t)(V_{os3} - V_{os4}) \quad (23)$$

$$V_{out2} = \frac{R_5}{R_3}\left(1 + 2\frac{R_1}{R_G}\right)\{V_{in}\sin(\omega_{in}t) + 2g_2(t)g_3(t)V_{os1} - 2g_2(t)g_3(t)V_{os2}\} + \frac{R_5}{R_3}g_2(t)(V_{os3} - V_{os4}) \quad (24)$$

From Eqs. 23 and 24, V_{in} is modulated and demodulated twice by Chopper1 to 4. In the meantime, V_{os1} and V_{os2} are modulated to a higher frequency band by Chopper3 and Chopper4. Furthermore, V_{os3} and V_{os4} are modulated to high frequency band by Chopper4. Figure 6 shows the flow of the input signals and the offset voltages in the second block.

As the results, we can discriminate the input signals and all of the offset voltages by using low-pass filter as shown in Figure 7.

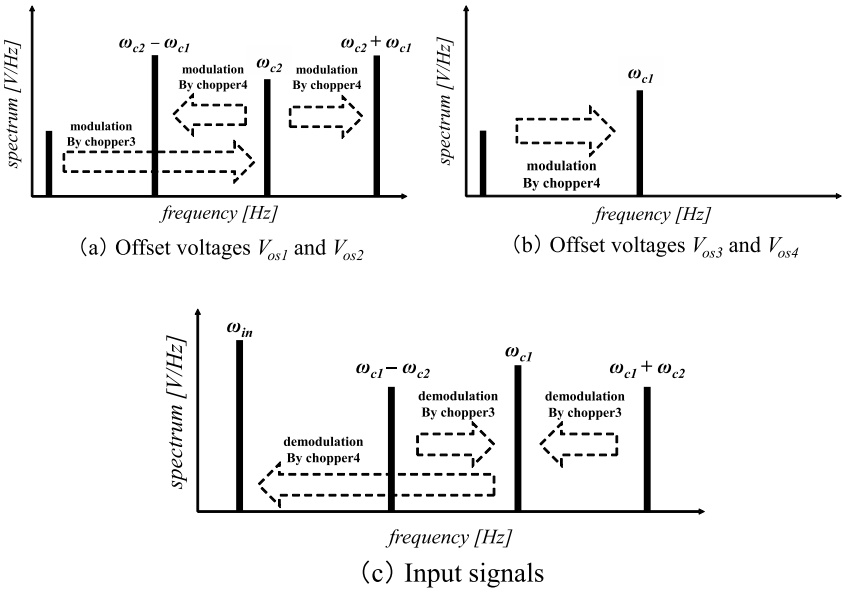


Fig. 6 The flow of the input signals and the offset voltages in the second block

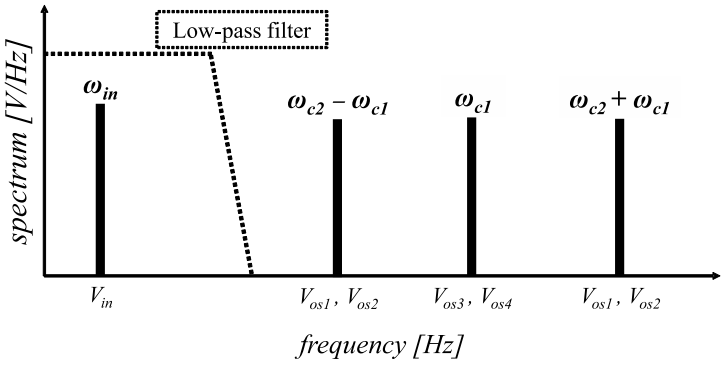


Fig. 7 The discrimination between the input signals and the offset voltages

4 Simulation Results

The performance of the proposed INA was evaluated by using HSPICE simulations. In this simulation, a set of parameters of 1P 2M 0.6μm CMOS is used. Figures 8 and 9 show the circuit schematics of Op-amps in the first and second blocks, respectively. The Op-amp in the second block has the differential output as shown in Figures 4 and 9. All MOSFETs except the output stage are operated in the subthreshold region in order to achieve the low power consumption. Since the load drivability is required in the output stage, the MOSFETs in the output stage are operated in the strong inversion region. Table 1 shows the conditions of these simulations.

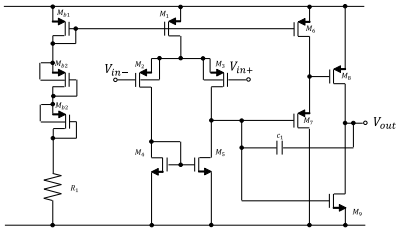


Fig. 8 The circuit schematic of Op-amp1 and Op-amp2

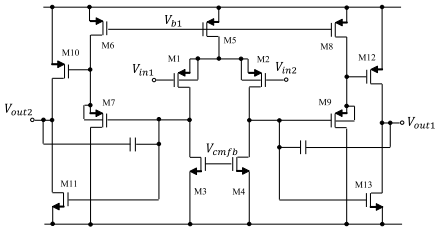


Fig. 9 The circuit schematic of Op-amp3

Table 1 The conditions of these simulations

Item	Value
Power supply[V]	$\pm$1.5 [V]
Amplitude of the input signals	0.3 [V]
Frequency of the input signals	100 [Hz]
Chopper frequency for SW1 (ω_{c1})	4 [kHz]
Chopper frequency for SW2 (ω_{c2})	10 [kHz]
V_{os1}, V_{os3}	25 [mV]
V_{os2}, V_{os4}	$-$25 [mV]

Figures 10 to 15 show the simulation results on each node. From these figures, we can confirm that the input signals and the offset voltages are modulated as theory. Furthermore, we can confirm that the final output signals after the low pass filter are amplified without affecting the offset voltage as shown in Fig. 15.

Figures 16 to 21 show the simulation results of FFT analysis on each node. From these figures, we could confirm that the input signals and the offset voltages are correctly modulated and demodulated by chopper switches as theoretical analysis.

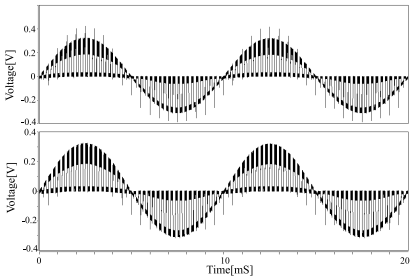


Fig. 10 Tansient analysis of V_a and V_b

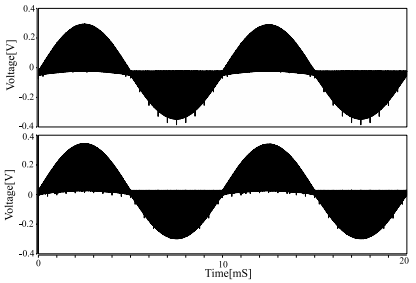


Fig. 11 Tansient analysis of V_c and V_d

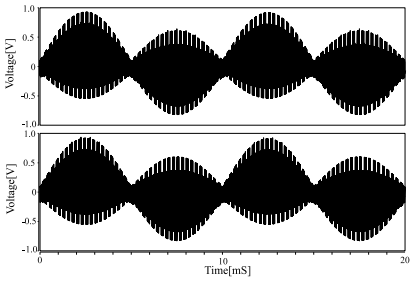


Fig. 12 Tansient analysis of V_e and V_f

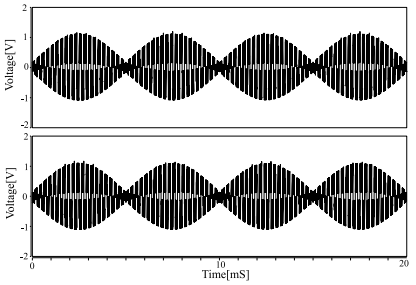


Fig. 13 Tansient analysis of V_g and V_h

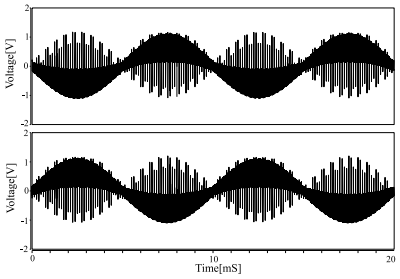


Fig. 14 Tansient analysis of V_{out1} and V_{out2}

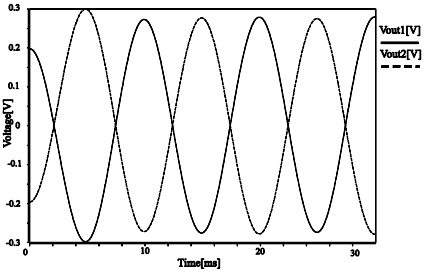


Fig. 15 Using the low-pass filter after V_{out1} and V_{out2}

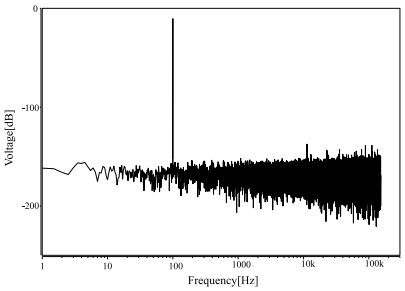


Fig. 16 FFT of the input voltage

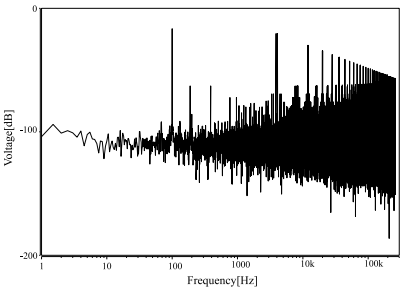


Fig. 17 FFT analysis of V_a and V_b

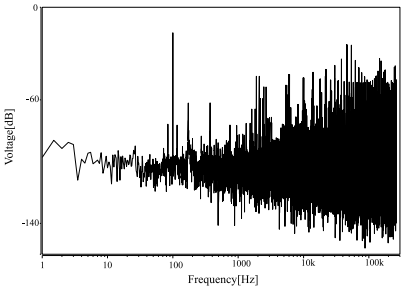


Fig. 18 FFT of V_c and V_d

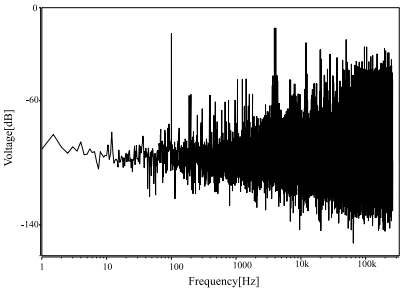


Fig. 19 FFT of V_e and V_f

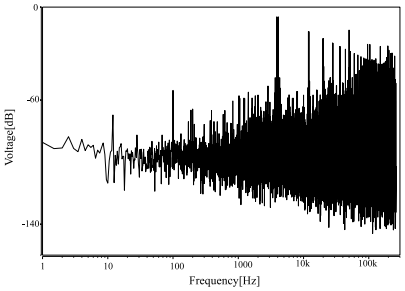


Fig. 20 FFT of V_g and V_h

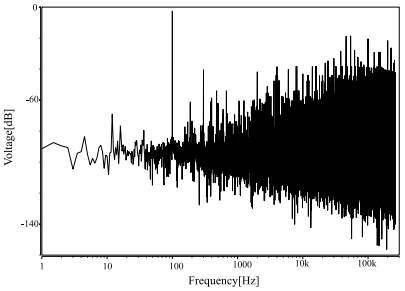


Fig. 21 FFT of V_{out1} and V_{out2}

5 Conclusion

In this paper, we have proposed the low offset voltage INA with the double chopper stabilization technique. Since the offset voltages can be perfectly modulated to high frequency band, the proposed circuit can remove the offset voltages by adding low-pass filter. The performance has been evaluated by using HSPICE simulation. As a result, we could confirm that double chopper stabilization technique contributes to remove the offset voltages.

The IC fabrication of the proposed circuit using this technique and its evaluation are future work.

References

1. Ueno, K., Horose, T., Asai, T., Amemiya, Y.: An Ultra-low Power Voltage Reference Circuit consisting of Subthreshold MOSFETs. IEEE Journal of Solid-State Circuits **44**(7), 2047–2054 (2009)
2. Ueno, K., Asai, T., Amemiya, Y.: Low-power Temperature-to-frequency Converter Consisting of Subthreshold CMOS Circuits for Integrated Smart Temperature Sensors. Sensors and Actuators A: Physical **165**(1), 132–137 (2011)
3. Nishimura, K.: Study on Instrumentation Amplifier Using Chopper Stabilization Technique. Master Thesis of the Faculty of Engineering, University of Miyazaki (2010)
4. Bakker, A., Thiele, K., Huijsing, J.: A CMOS Nested-Chopper Instrumentation Amplifier with 100-nV Offset. IEEE Journal of Solid-State Circuits **35**(12), 1877–1883 (2000)
5. Iwaki, M., Nakashima, Y., Toyama, T., Tanno, K., Ishizuka, O.: A Study on Low-Voltage, Low-Power and Low-Noise Amplifier for Surface-Myoelectricity Signal Processing LSI. Memoirs of the Faculty of Engineering, Miyazaki University **33**, 187–193 (2004)
6. Tamura, H., Gotoh, T., Okumura, D., Tanaka, H., Tanno, K.: A study on the s-EMG pattern recognition using neural network. In: International Symposium on Intelligent Informatics (2008)
7. Tamura, H., Tanno, K., Tanaka, H., Catherine, K., Zheng, T.: Recurrent type ANFIS using local search technique for time series prediction. In: IEEE Asia Pacific Conference on Circuits and Systems, Macao, Japan (2008)
8. Razavy, B.: Design of Analog CMOS Integrated Circuits. McGraw-Hill (2000)
9. Gray, P.R., Hurst, P.J., Lewis, S.H., Meyer, R.G.: Analysis and Design of Analog Integrated Circuits. Jhon Wiley & Sons (2001)
10. Ueno, K.: CMOS voltage and current reference circuits consisting of subthreshold MOSFETs -micropower circuit components for power-aware LSI applications. In: Swart, J.W. (ed.) Tech. Solid State Circuits Technologies, pp. 1–24 (2010)

A Study on Human Interface for Communication Using Electrooculogram Signals

Kazuya Gondou, Hiroki Tamura and Koichi Tanno

Abstract Human interface using eyes for a person with disabilities has been researched. Almost ALS (Amyotrophic Lateral Sclerosis) patient can move facial muscle and eyeball. Therefore, human interface using eyes is the communication tool for a person with disabilities. Human interface is very important when aiming at improvement of quality of life. There are various gaze recognition methods. For example, camera [1] and search coil method [2] are general techniques. In this paper, we propose a human interface for communication using electrooculogram method by 4 electrodes. From the simulation results, our system has high accuracy of eyes pattern classification.

Keywords EOG · Electrooculogram signal · Drift · Electromyogram signals

1 Introduction

Patients of ALS in Japan reached 20,000 people by an investigation in 2012 [3]. ALS is the neurosis to get exercise is infringed on, and movement failures occur by numbness of hand and foot. Moreover a respiratory muscle weakens, breathing become difficult. When the limb function abolishes finally, utterance have also become difficult, the expression means by their own effort will be lost [4]. However, failures don't occur in the five senses generally because a sensory nerve and an independent nerve aren't infringed in ALS. Moreover, failures about eyeball movement rarely occur. Therefore, human interface using eyeballs is studied extensively as the means of the effective expression for ALS patient [1-2,4-6].

K. Gondou(✉) · K. Tanno
Department of Electrical and Electronic Engineering,
University of Miyazaki, Miyazaki, Japan
e-mail: tc14008@student.miyazaki-u.ac.jp

H. Tamura
Department of Environmental Robotics, University of Miyazaki, Miyazaki, Japan

T.T. Zin et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 388,
DOI: 10.1007/978-3-319-23207-2_31

Also human interface using eyeballs by electrooculogram (EOG) is studied by many researchers. EOG method is a method to recognition a gaze using the change in feeble EOG when sticking electrodes on a face and changing a gaze [5]. However, EOG method had the drift problem. The Drift is that the direct current element of EOG changing every moment by the time course. Conventional EOG method needs five electrodes (four electrodes on face and one grand electrode).

In this paper, we proposed a novel EOG method specializing in horizontal direction using four electrodes. The electrodes of our proposal method are one less than the conventional method. However, our proposal method also has the technique to reduce the drift. The validity of our proposal method is compared with conventional method by the point of eyes pattern classification and the drift problem.

2 Conventional Method

By a conventional method [7], DC (Direct Current) and AC (Alternate Current) of EOG acquired by five electrodes attached as shown in Fig.1. Gaze and movement of facial muscle are recognized at the same time. To arrange electrodes, waveform with the feature different getting a finite difference of 2 channels (ch1, ch2) and getting a total is obtained. Waveform when changing a look in left, right, up or down, is shown in Fig.2. The vertical axis shows amplitude of EOG and the horizontal axis is time. DC1 is DC of ch1, DC2 is DC of ch2. Getting a finite difference of EOG 2 channels (DC1-DC2), the amplitude of when changing a look in the left and right direction is emphasized. And getting the summation of 2 of EOG, the amplitude when changing a look in the top and the bottom, is emphasized. Using this feature, 4 directions eye movement is distinguished. Algorithm of a conventional method is shown in Fig.3. Fig.3 is flowchart of the conventional method from which both of surface electromyogram (sEMG) and EOG are distinguished at the same time. When DC and AC elements exceeded a fixed threshold at the same time, gaze is judged to have moved to one of them in left, right, up or

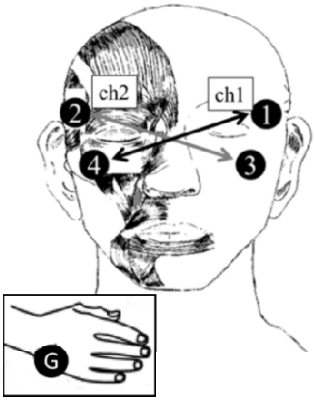


Fig. 1 Electrode position by conventional method ("G" is a ground)

down. When gaze faces to the center and AC is stable, the central value which becomes the standard is renewal. In this way, three patterns of sEMG movement bite, right blink and left blink, and four patterns of gaze movement left, right, up and down can be recognized.

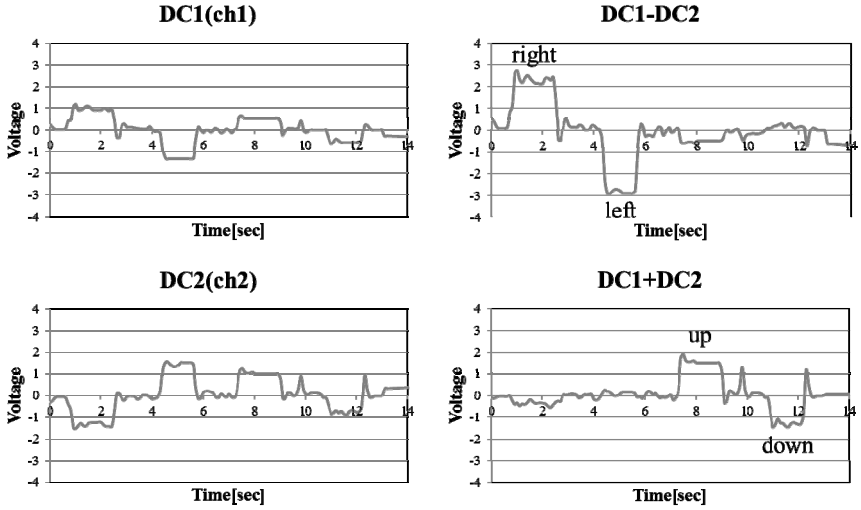


Fig. 2 Waveform when eye move 4 directions

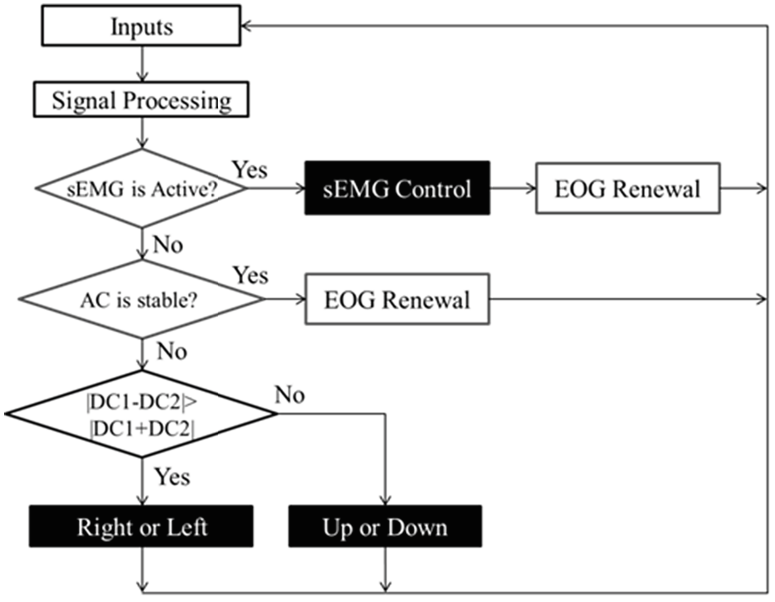


Fig. 3 Algorithm of the conventional method

The experimental equipment is the human machine interface developed by ours for persons with disabilities. Our human interface is shown in Fig.4. Our human machine interface is possible to control PC and electric wheelchair using the sEMG and EOG. Gaze movement using our human machine interface in left and right can be distinguished by the precision of 99.1%. However, there is a problem that a drift isn't canceled perfectly yet. And to use five electrodes, there is a trouble in an attachment sense. Therefore, we propose the novel EOG method try to solve these problems.

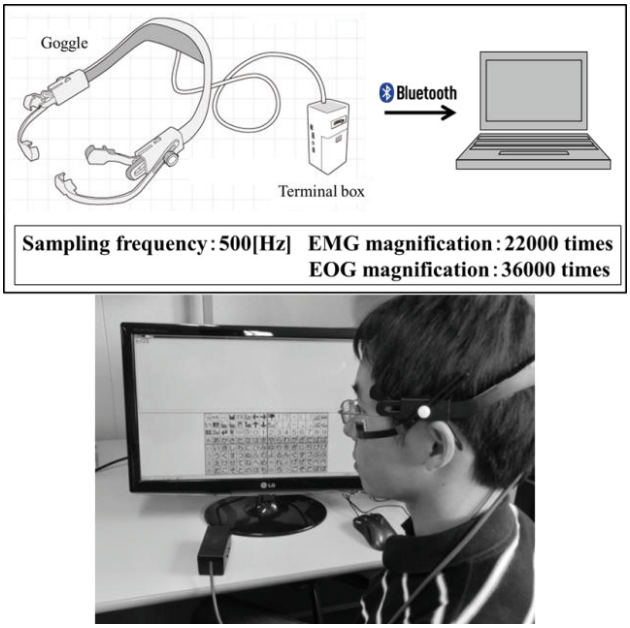


Fig. 4 Human interface device

3 Proposed Method 1

First, the ground stuck on the arm is moved to the face and integrate the underside electrode into one electrode (electrode “3” in Fig.5), and simply the number of electrode is set to only 4 attached to a face. An attachment figure of electrode is shown in Fig.5. In this state, a gaze recognition experiments in the left and right direction were tried.

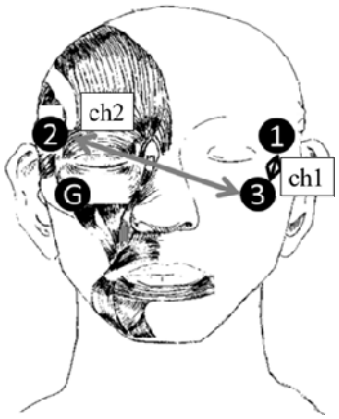


Fig. 5 Electrode position by proposed method ("G" is a ground)

3.1 Experiments and Results

Experiments are that character input by our human machine interface and 2 patterns mode of the software called "Heartyladder". Heartyladder is character input support software for persons with disabilities. The way of character input is shown in Fig. 6. As shown in Fig. 6, when moving a gaze to the right, the right side of the dial is chosen, and when moving it to the left, the left side is chosen. The chosen reach of the dial is becoming small gradually by repeating these and one character is chosen finally. The characters from "A" to "Z" were inputted by using Heartyladder in this way, and the experiments which check the recognition rate of the case is made to one subject. As the results, the recognition rate of our proposed method using four electrodes have fallen into 95.3% compared with the conventional method 99.1% in left and right, so cause is investigated.

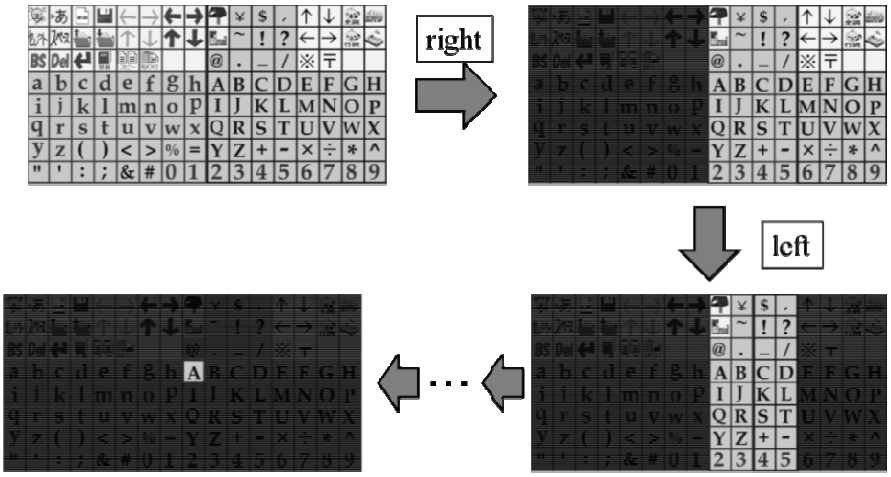


Fig. 6 Character input using Heartyladder

3.2 The Cause of the Recognition Rate Fall

The waveform of EOG of conventional method and proposed method is shown in Fig.7. The vertical axis shows the amplitude of EOG and the horizontal one is time. EOG is acquired from 2 channels like Fig.1 and Fig.5 and a finite difference of the value the 2 channels gave is being used for gaze recognition in the left and right direction. As shown in Fig.7, when turning a gaze to left and right by the conventional method, DC1 and DC2 are changing into an opposite direction strongly, and the thing for which these 2 finite differences could give us a strong amplitude of vibration. On the other hand getting these 2 finite differences, the amplitude of vibration have been rather small because DC1 and DC2 are changing into the same way by proposed method as shown in Fig.8. Therefore the threshold value is difficult for the change amount of EMG to exceed more than the conventional method, and the precision has fallen. Even if the summation of EOG of 2 channels is used for recognition in left and right, it won't be effective solution method because the change amount of DC1 is very small.

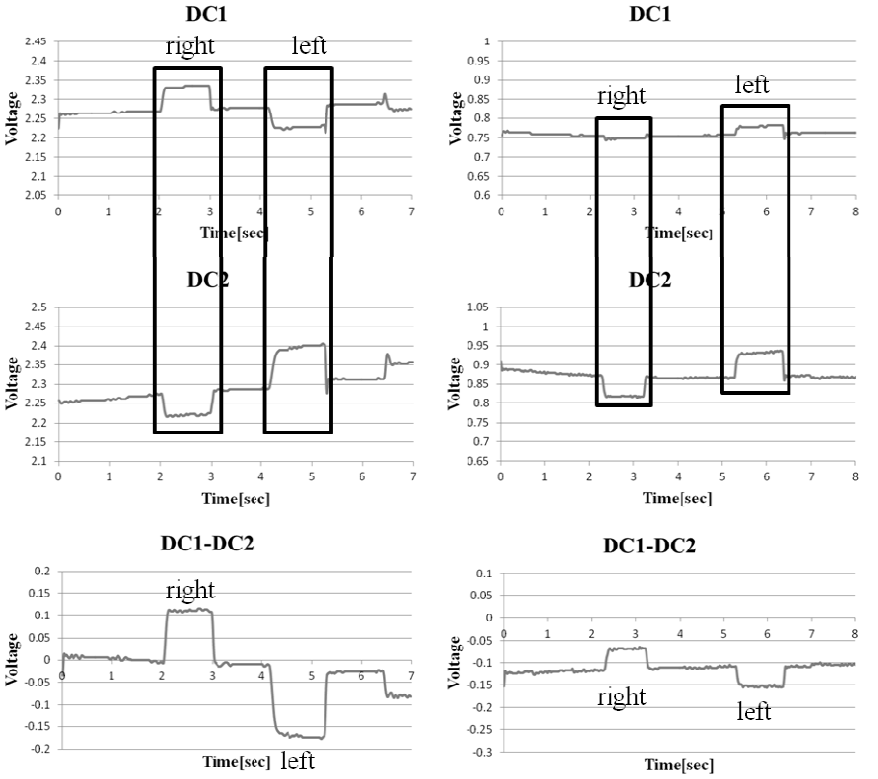


Fig. 7 Comparison with the conventional method (left) and the proposed method (right)

4 Proposed Method 2

To improve the problem, electrode position is changed. New electrode position is shown in Fig.8. The underside electrode is brought close to a facial center, near the nose. EOG waveform near a conventional method might be obtained from purpose in which eyes are located between electrode "1" and "3" or electrode "2" and "3" by arranging this. Moreover, one of the causes the drift is the bloodstream which flows through the skin surface [8], and there is a possibility that a drift can be reduced by arranging a pole around a little nasal bone of the blood flow volume. Even if a drift occurs, the reference electric potential of 2 channels is common, and DC transfers in the same direction. Therefore the drift is offset when getting a finite difference of 2 channels of DC.

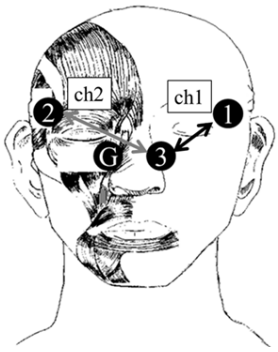


Fig. 8 Electrode position by the proposed method 2

5 Experiments and Results

5.1 Recognition Rate

When the same experiment as section 4 is conducted to 2 subjects by new electrode position, the precision equal to recognition rate 99.1% and a conventional method's 99.1% mostly is obtained. EOG waveform by new electrode position is shown in Fig.9. The vertical axis shows amplitude of EOG and the horizontal axis is time. From Fig.9, DC1 and DC2 are changing into an opposite direction at the time of a gaze movement in left and right like the conventional method. The enough amplitude of vibration is obtained by proposed method 2, therefore high precision like the conventional method is obtained.

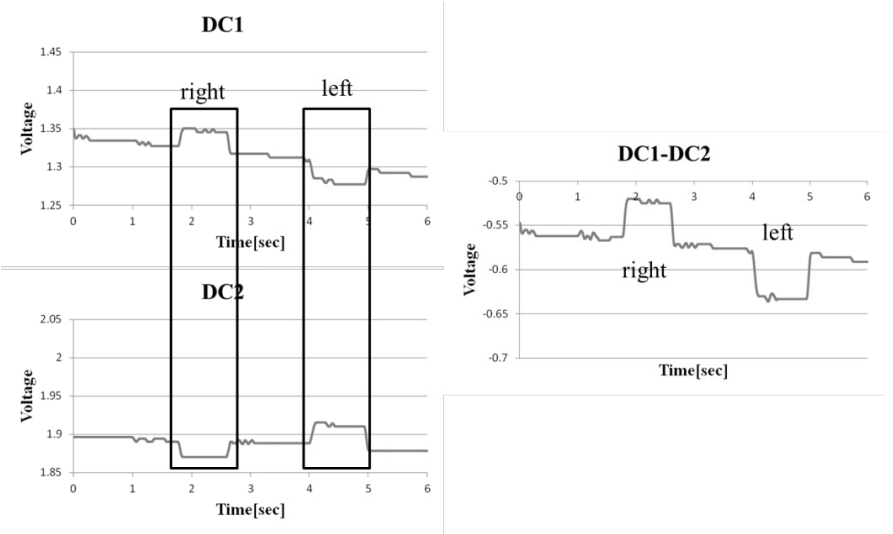


Fig. 9 EOG by the proposed method 2

5.2 Reduction in the Drift

In this subsection, we investigate about the drift. The change in the drift of DC when keeping seeing the center of the PC screen for 30 minutes, is measured in both of a conventional method and improved proposed method. This experiment is also conducted to 2 subjects. Experimental result is shown in Fig.10 and Fig.11. The vertical axis of Fig.10 shows amplitude of EOG and the horizontal axis is

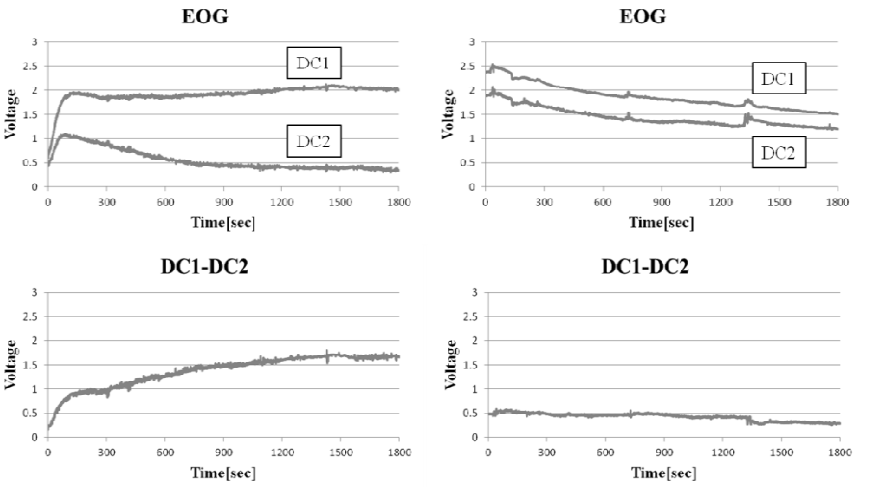


Fig. 10 Comparison of the drift by the conventional method (left) and the proposed method 2 (right)

time. As shown in Fig.10 left, getting a finite difference of 2 channels, it increased in influence of the drift because the way of the drift is different in DC1 and DC2 by the conventional method. On the other hand, and the proposed method 2 shows that influence of the drift getting a finite difference, is reduced because DC1 and DC2 are transferring in the same direction as shown in Fig.10 right. Fig.11 indicates when the drift amount average of 2 subjects in DC1-DC2 is compared by 2 methods. Proposed method 2 shows that the drift amount from measurement starting to the end is reduced 29.8 % compared with a conventional method, and it's reduced 51.0 % about the maximum variation amount of the drift which can be put during measurement.

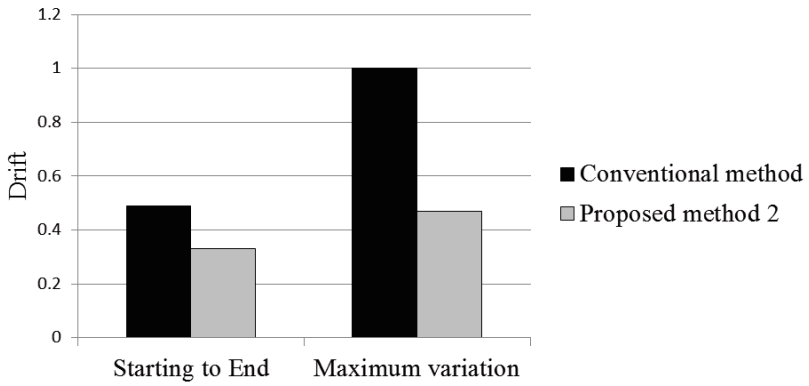


Fig. 11 Comparison of the drift amount

6 Conclusion

In this paper, we introduced the gaze recognition by the number of four electrodes. Form these results, our proposed method had recognition precision equal to conventional method using five electrodes. Moreover, the drift reduction succeeded in getting a finite difference of 2 channels. As mentioned previously, the value of a finite difference of 2 channels was taken is used for gaze recognition in left and right. Therefore, gaze recognition in the more precise left and right direction is expected by using our proposed method. There is a possibility that not only gaze direction but also seen area is can be judged. On the other hand, proposed method has a fault that the upper and lower recognition becomes more difficult because this method is specialized in recognition of the left and right. Therefore, the upper and lower recognition rate improvement can be presented as future's problem.

References

1. Kadyrov, A., Yu, H., Eyles, J., Liu, H.: Explore new eye tracking and gaze locating methods. In: 2013 IEEE International Conference on Systems, Man, and Cybernetics, pp. 2866–2871 (2013)
2. Houben, M.M.J., Goumans, J., van der Steen, J.: Recording Three-Dimensional Eye Movements: Scleral Search Coils versus Video Oculography. Investigative Ophthalmology & Visual Science **47**, 179–187 (2006)
3. Annual Report on Government Measures for Persons with Disabilities (Summary) (2012)
4. Yagi, T., Miyanaga, A., Numata, H., Koga, K., Funase, A., Mukai, T.: A Study on an Eye-gaze Interface using a Biological Signal. JSME Annual Meeting **2006**(5), 575–576 (2006)
5. Kuno, Y., Yagi, T., Fujii, I., Koga, K., Uchikawa, Y.: Deveolpment of Eye-gaze Input Interface using EOG. Transactions of Information Processing Society of Japan **39**(5), 1455–1462 (1998)
6. Manabe, H., Fukumoto, M.: Full-time wearable headphone-type gaze detector. In: Proceeding CHI EA 2006 CHI 2006 Extended Abstracts on Human Factors in Computing Systems, pp. 1073–1078 (2006)
7. Yan, M., Tamura, H., Tanno, K.: Gaze estimation using electrooculogram signals and its mathematical modeling. In: Proceedings CDROM, IEEE 43rd International Symposium on Multiple-Valued Logic 2013, May 21-24, 2013, Toyama, Japan, pp. 18–22 (2013)
8. Yanagimoto, T., Maruki, K., Ueno, E., Wang, G., Yoshida, H., Yunokuchi, K.: A Study on Autonomic Nervous System by Analysis of Skin Potential Level. IEICE Technical Report. ME and Bio Cybernetics **100**(599), 13–19 (2001). (Japanese)

A Study on Indoor Presence Management System Using Smartphone

Takami Taninoki, Yoshinobu Furukawa, Hiroaki Matsumoto, Hiroki Tamura and Koichi Tanno

Abstract The aim of this paper is to perform indoor presence management using smartphone. Our proposed method estimates user position and state by smartphone sensors. The position information is estimated by accelerometer and direction sensor. The state information is estimated using gravitational acceleration can be acquired by accelerometer. By using this system, it can estimate human position within 100 cm errors. In addition, this system can estimate the state of subject in 99.2 % accuracy.

Keywords Presence management · Smartphone · Position estimate method · State estimate method

1 Introduction

In recent years, study on presence management has been becoming actively. The presence management is to manage information of position and state of people or things. For example, you can change the contact methods in accordance with the partner state in the scene of business, it is possible by using the presence management. In addition, it can be expected to use as a watch system which can manage

T. Taninoki(✉) · Y. Furukawa · H. Matsumoto
Department of Electrical and Electronic Engineering,
University of Miyazaki, Miyazaki, Japan
e-mail: tc14016@student.miyazaki-u.ac.jp

H. Tamura
Department of Environmental Robotics, University of Miyazaki, Miyazaki, Japan

K. Tanno
Department of Electrical and Systems Engineering,
University of Miyazaki, Miyazaki, Japan

T.T. Zin et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 388,
DOI: 10.1007/978-3-319-23207-2_32

the position and state of the indoor people. Next, we explain the characteristic of various sensors that is used in research on the presence management.

1. The method using visual sensor[1][2]

Visual sensor can get a lot of information not only the position of subject but also appearance and gestures if using the technology of image recognition. In many cases visual sensor can also personal identification at the same time. However, visual sensor has disadvantage that affected by the lighting conditions and lose visual contact for the subject by occlusion. In addition, the problem of privacy is relatively large. Because, visual sensor can recognize the appearance.

2. The method using laser range scanner[3][4]

Laser range scanner measures the distance to subject using reflected light when irradiating the subject with laser. Therefore, high accuracy position estimation can be expected because laser range scanner is not affected much by the irradiation conditions. However, personal identification is difficult because laser range scanner can't get much information. In addition, laser range scanner also can be affected by occlusion as same as visual sensor.

3. The method using portable sensor[4]

Portable sensor exist ID tag and acceleration sensor as example. Personal recognition and position estimation can be enabled to the subject who has these devices always. Therefore, personal recognition almost certainly enabled by getting the ID information. However portable sensors are low accuracy.

Above sentence shows the various techniques for realizing the presence management. From the point of privacy and occlusion problem, we select the portable sensor in this paper. We show our proposed system and experimental results of position estimation and state estimation. In the position estimation, subject moves the route of 5 patterns at indoor. In the state estimation, subject does the state of 4 patterns at indoor. Lastly, we show the accuracy of position estimation and state estimation. In the conclusion, we compare with other researches on the presence management.

2 Proposed Method

In this section, we show the explanation of proposed system.

2.1 Position Estimation Method

Position estimation uses the data of acceleration sensor and direction sensor when subject holds smartphone by hand. Position estimation uses the data of acceleration sensor and direction sensor, when subject is moving. Walking or stopping is

judging from the data of acceleration. In the case of walking, coordinate of position is calculated by estimate equation. Fig.1 shows the route of preliminary experiment. Fig. 2 shows the orientation of the acceleration sensor.

First, we explain the method of walking judgement. The z-axis data of Fig. 3 found to have changed significantly during walking. Vertical axis in Fig. 3 is the value of acceleration. The threshold of walking judgment is -0.15. Our proposed method determines "walking" when the z-axis acceleration is less than -0.15. Proposed method needs determining of "walking" while the z-axis acceleration is fluctuating periodically. Therefore, proposed method determines "walking" during 0.6 seconds from the z-axis acceleration is lower than threshold. Next, we explain how

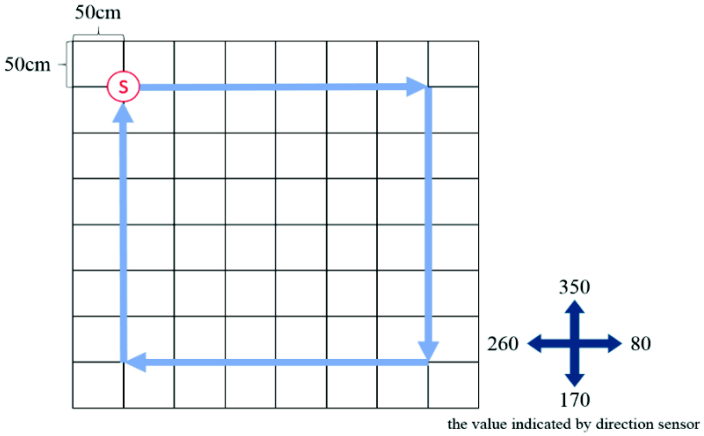


Fig. 1 Route of preliminary experiment

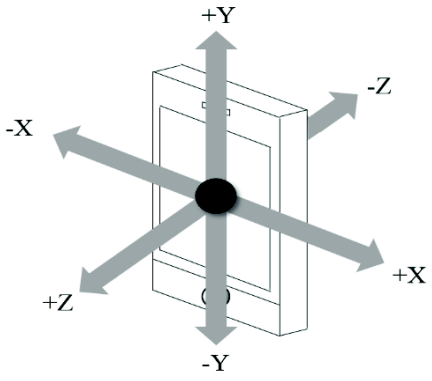


Fig. 2 The orientation of the acceleration sensor

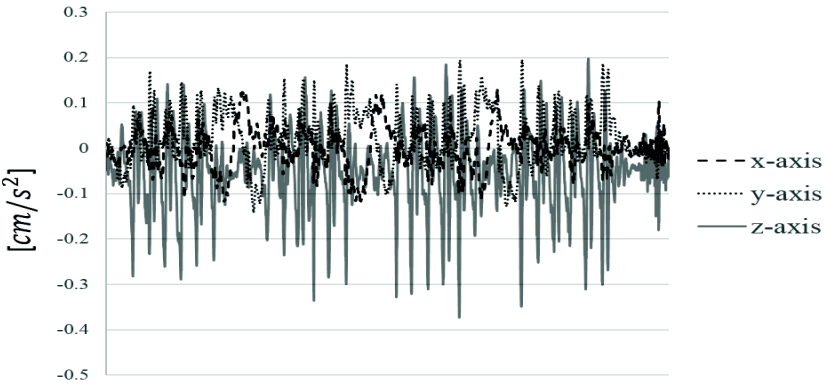


Fig. 3 The data of acceleration sensor when preliminary experiment

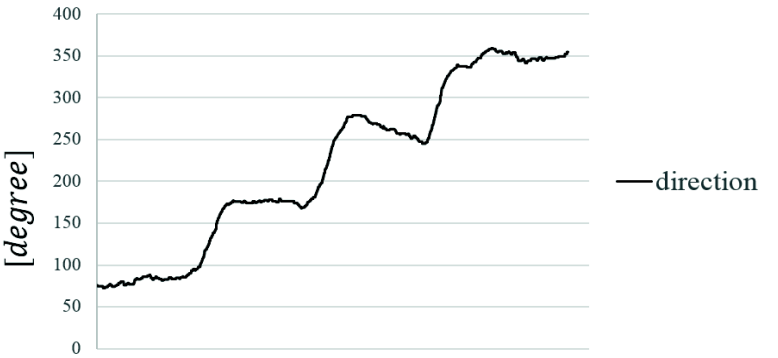


Fig. 4 The data of direction sensor when preliminary experiment

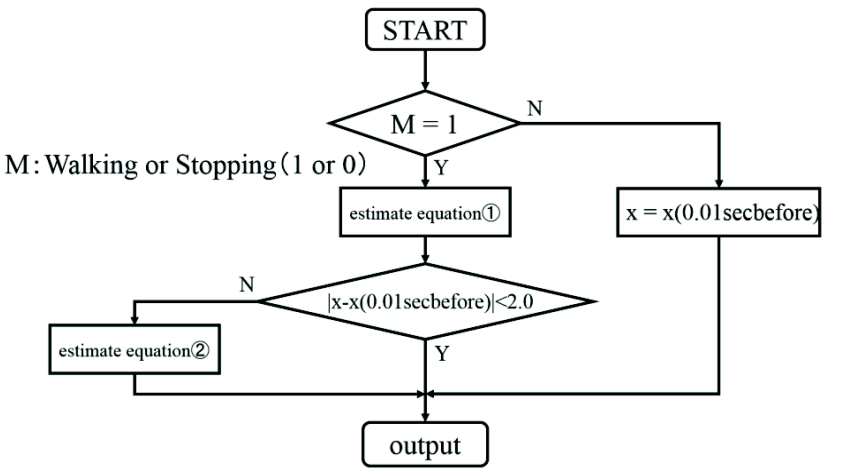


Fig. 5 The flowchart of the position estimation

to handle the data of direction sensor. Fig. 4 shows the data of direction sensor. Vertical axis in Fig. 4 is the value of direction sensor. And the right side of Fig. 1 shows the value indicated by direction sensor in experiment room. In addition, the position of subject is estimated by substituting data to equation of estimation.

$$posX(t) = v * \sin(dir)(t) * M(t) + posX(t-1) \quad (1)$$

$$posY(t) = -v * \cos(dir)(t) * M(t) + posY(t-1) \quad (2)$$

$$posX(t) = (v/100) * \sin(dir)(t) * M(t) + posX(t-0.01) \quad (3)$$

$$posY(t) = -(v/100) * \cos(dir)(t) * M(t) + posY(t-0.01) \quad (4)$$

Equation (1), (2), (3), and (4) shows estimate method of position. Characters "posX" and "posY" are the position of subject, "dir" is direction, "m" is the value of walking (=1) or stopping (=0). Estimating value is determined by multiplying the direction components and the walking speed, further add coordinate of 1 second ago or coordinate of 0.01 second ago. In this case the walking speed “v” is to be constant, this speed is set to 71cm per sec. Fig. 5 shows flowchart of the position estimate. Proposed method uses equation (1) and (2), or (3) and (4) while user is walking. Equation (1) and (2) estimates present position using coordinate of 1 second ago. Equation (3) and (4) are used when the difference between estimation value of equation (1) or (2) and 0.01 second ago coordinate becomes 2 or more. When not walking, output is coordinates of 0.01 second ago. The number 0.01 is used because sampling frequency of acceleration and direction are 100Hz.

2.2 State Estimation Method

State estimation acquires the data of gravitational acceleration when subject does various actions. In this time, there is subject smartphone in the pocket of trouser. There are 4 kinds of state estimation, standing, walking, sitting, and crouching. Algorithm of state estimation is considered by the characteristic of gravity acceleration.

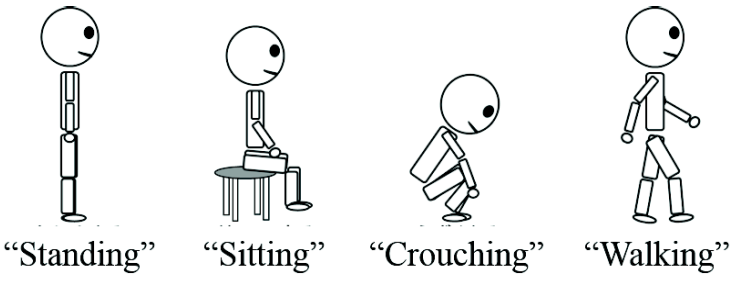


Fig. 6 The kind of states

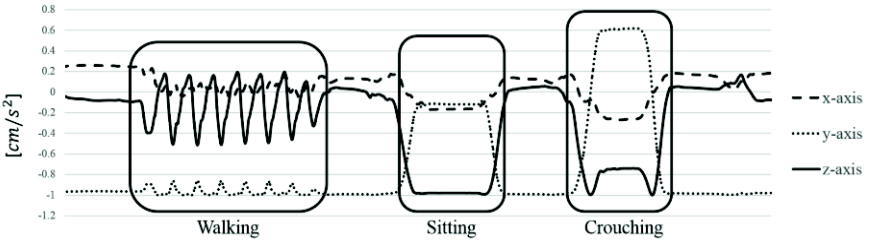


Fig. 7 Gravitational acceleration

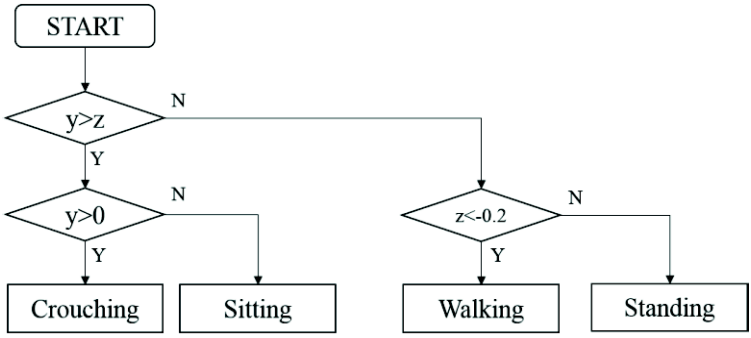


Fig. 8 The flowchart of the state estimation

Fig. 6 shows the kind of states. First, state of "standing" and "walking" are state-group 1. In addition, state of "sitting" and "crouching" are state-group 2. Fig. 7 shows gravitational acceleration that measured by smartphone in subject trouser pocket when subject does various actions. Vertical axis in Fig. 7 is the value of gravity acceleration. The point of attention is gravity acceleration of y-axis and z-axis. The value of z-axis is larger than value of y-axis in the case of state-group 1. On the other hand, the value of y-axis is larger than value of z-axis in the case of state-group 2. Therefore, these 2 groups can divide by conditional equation. In the next step, we explain how to judge "standing" and "walking". Focusing on the gravity acceleration of z-axis during walking, it can be seen that the values are changed periodically. We can estimate that “walking” when the gravity acceleration of z-axis is lower than the threshold value -0.2. Otherwise, we can estimate that "standing". In this case, the estimated state is needed the state of "walking" during the gravity acceleration of z-axis changes periodically. Therefore, state estimation system estimates "Walking" 0.6 seconds from the gravity acceleration of z-axis falls below the threshold. In the next step, we explain how to judge "sitting" and "crouching". Focusing on the gravitational acceleration of y-axis, the value of "crouching" is lower than the value of "sitting". Therefore, we can estimate that "sitting" when the gravitational acceleration of y-axis is lower than the threshold value 0. Otherwise, we can estimate that “crouching”. In addition, assuming that the state of "sitting" or "crouching" doesn't migrate directly to "walking", adds the state of "processing".

3 Experiments

3.1 Position Estimation Experiment

In order to examine the accuracy of position estimation by proposed method, subject walks in the room holding smartphone in hand. There are 5 walking route, and starting point is unified. In the position estimation, 4 estimate equations are used. Therefore, we show the difference between equation (1), (2) and equation (3), (4). The experiment results shows below.

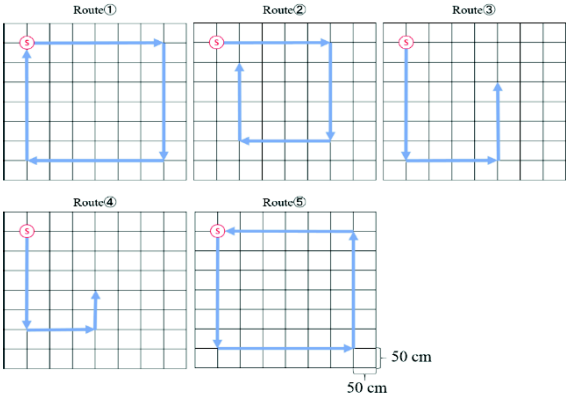


Fig. 9 The walking route of experiment

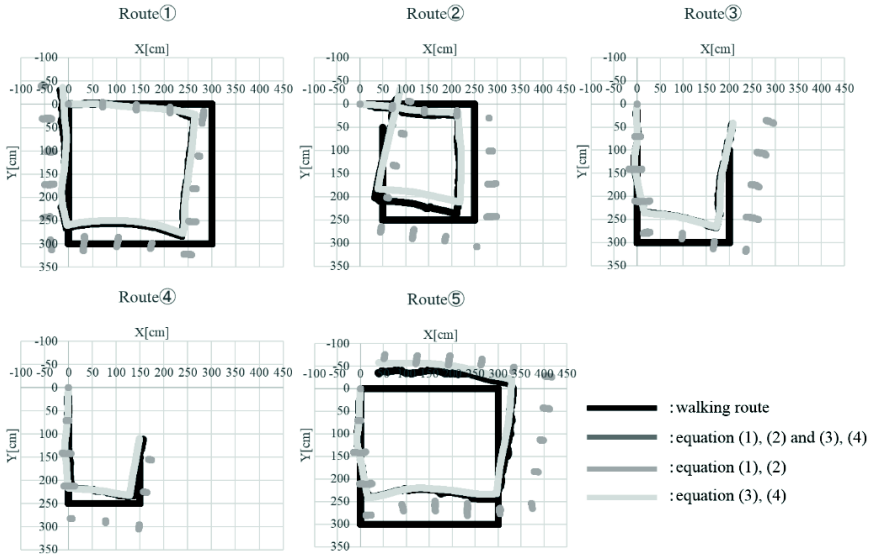


Fig. 10 Estimated results of position

Table 1 Accuracy of equation (1), (2)

Equation (1), (2)	RMSE X[cm]	RMSE Y[cm]	Maximum error X[cm]	Maximun error Y[cm]
Route①	46.2	33.0	125.5	110.2
Route②	28.6	36.7	70.1	100.0
Route③	42.3	30.1	97.5	79.0
Route④	18.9	35.5	77.2	69.9
Route⑤	71.6	46.7	117.6	115.0
Average	41.5	36.4	97.6	94.8

Table 2 Accuracy of equation (3), (4)

Equation (3), (4)	RMSE X[cm]	RMSE Y[cm]	Maximun error X[cm]	Maximun error Y[cm]
Route①	33.9	30.0	60.7	50.7
Route②	25.7	48.0	38.1	82.2
Route③	15.1	46.3	28.2	63.7
Route④	12.2	25.6	22.0	40.4
Route⑤	21.0	52.2	39.8	80.0
Average	21.6	40.4	37.8	63.4

Table 3 Accuracy of equation (1), (2) and (3), (4)

Equation (1), (2) and (3),(4)	RMSE X[cm]	RMSE Y[cm]	Maximun error X[cm]	Maximun error Y[cm]
Route①	35.2	25.8	63.8	44.4
Route②	27.1	31.4	39.4	63.6
Route③	14.8	43.5	28.4	65.2
Route④	7.6	23.5	14.0	38.9
Route⑤	20.2	43.5	38.0	78.9
Average	21.0	33.6	36.7	58.2

Fig. 9 shows the walking route of experiment. Fig.10 shows estimated result of position. In the Fig.10, vertical axis is X[cm] and horizontal axis is Y[cm]. As shown in table 1, table 2, and table 3, estimate equation (3), (4) has higher accuracy when comparing the results of the estimate equation (1), (2) and estimate equation (3), (4). However, sometimes the estimate equation (1), (2) has true value nearer than the value of estimate equation (3), (4). Therefore, we compares the result of each estimation equation. Estimate equation (3), (4) has higher accuracy sometimes but estimate equation (1), (2) and (3), (4) has higher accuracy furthermore on the whole. Therefore, proposed estimate equation can estimate position that has error less than 100 cm.

3.2 State Estimation Experiment

In order to examine the accuracy of state estimation by proposed method, three subjects acted in the room that has smartphone in pocket of trouser. The subjects try each states in steps of 20 times, and examines those success rate. We show the experimental results at below.

Fig.11 shows the result of state estimation when subject does state changing. Vertical axis in Fig.11 is the value of state estimation. First, the result of state

estimation is "Walking" when the state of "Standing" just changes "Crouching" or "Sitting". It can be seen from the Fig.7, because condition of "Walking" was satisfied when actions shift. In addition, the state of subject go through "Sitting" from human nature when changing the state of "Crouching" from "Standing". It is preferable those can be changed to "Processing". On the other hand, the state estimation judged "Processing" when the state of "Crouching" or "Sitting" changed from "Standing". Next, table 4 shows the success rate of state estimation. The "Walking" result of two subjects became 95%. That's because slow tempo of walking and gravity acceleration doesn't over threshold within 0.6 second. The average of all success rate was 99.2%. Therefore, this system can perform highly accurate state estimation which is needed for state management.

Fig. 11 The result of state estimation

Table 4 Success rate of state estimation

	Subject A	Subject B	Subject C	Average
Standing	100[%] (20/20)	100[%] (20/20)	100[%] (20/20)	100[%]
Walking	100[%] (20/20)	95[%] (19/20)	95[%] (19/20)	96.7[%]
Crouching	100[%] (20/20)	100[%] (20/20)	100[%] (20/20)	100[%]
Sitting	100[%] (20/20)	100[%] (20/20)	100[%] (20/20)	100[%]
Average	100[%]	98.8[%]	98.8[%]	99.2[%]

4 Conclusions

The aim of this paper was to perform indoor presence management using smartphone. Therefore, experiments were conducted by building state management system consisting of position estimation and state estimation.

Table 5 Comparison of presence management using various sensors

	Accuracy[mm]	Personal recognizability	Privacy protection
Visual sensor	10^2〜$5*10^2$	△	×
Laser range scanner	10^2〜$5*10^2$	×	○
Portable sensor	10^3	○	○
Proposed method	10^2〜10^3	○	○

Table 5 shows the comparisons of our proposed method with other sensors. In the proposed method, privacy and occlusion problems have almost no effect. Because we used smartphone. The accuracy of proposed position estimation is inferior to the visual sensor and laser range scanner. However, smartphone can estimation of position always because it has no occlusion. By using proposed method, position estimation is less than 100 cm accuracy and state estimation is almost possible 100% accuracy. Using ID tags and improving accuracy of the position estimation within 50 cm will be implemented in the future. Our future work is to do the presence management by wearing smartphone everywhere.

References

1. Yu, G., Hu, Z., Lu, H., Li, W.: Rubust object tracking with occlusion handle. Neural Comput. & Applic., 1027–1034, June 2010
2. Black, M.J., Jepson, A.D.: EigenTracking: Robust Matching and Tracking of Articulated Objects Using a View-Based Representation. International Journal of Computer Vision **26**(1), 63–84 (1998)
3. Glas, D.F., Miyashita, T., Ishiguro, H., Hagita, N.: Laser tracking of human body motion using adaptive shape modeling. In: IEEE/RSJ International Confidence on Intelligent Robots and Systems, pp. 603–608 (2007)
4. Shibutani, S., Tamura, H., Tanno, K.: Human tracking and estimation under occlusion using laser range scanner and accelerometer. In: IEEE International Conference on Systems, Man, and Cybernetics, pp. 4196–4200 (2013)

A Study on sEMG Pattern Classification Method of Muscles of Respiration

Ryosuke Kokubo, Shogo Okazaki, Misaki Shoitizono, Hiroki Tamura and Koichi Tanno

Abstract The aim of this paper studies the possibility of new method to diagnose the sleep apnea syndrome. In this paper, we propose analysis method for the pattern classification of breathing from surface electromyogram. First, we measure surface electromyogram that obtained from the surface electrodes attached to crest of neck and mandible muscles. Next, we obtain the peak signal of active from Wavelet transformation of surface electromyogram. We calculate the pattern classification by using the k-nearest neighbor method. From the experimental results, our analysis method was possible to obtain high pattern classification rate when k is 6.

Keywords: Sleep apnea syndrome · sEMG [1]

1 Introduction

Although illness related to breathing is more than one, it has been noted in particular recent sleep apnea syndrome. Sleep apnea syndrome, is a disease that is to stop breathing or low breathing during sleep. Until now, when performing precision inspection, it has been determined by finally doctors and specialized clinical laboratory technologist from a plurality of test items that were admitted to the hospital.

R. Kokubo(✉) · S. Okazaki
Department of Electrical and Electronic Engineering, University of Miyazaki, Gakuenkibanadainishi 1-1, Miyazaki-City, Miyazaki, Japan
e-mail: tc14007@student.miyazaki-u.ac.jp

M. Shoitizono · H. Tamura
Department of Environmental Robotics Engineering, University of Miyazaki, Gakuenkibanadainishi 1-1, Miyazaki-City, Miyazaki, Japan

K. Tanno
Department of Electrical and System Engineering, University of Miyazaki, Gakuenkibanadainishi 1-1, Miyazaki-City, Miyazaki, Japan

T.T. Zin et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 388,
DOI: 10.1007/978-3-319-23207-2_33

In this paper, we propose method to classify the respiration and other operations. In future application, it make a diagnosis and treatment of simple manner without stress by possible to know the myoelectric potentials of the breathing.

2 Experiment

The experiments attached an electrode in two locations to near geniohyoid under the chin and near the scalene muscle of the neck. In this case, by using the skin pure to remove the horny body surface noise is prevented ride. We start the experiment from the electrode attached at about 10 minute time. Further, we confirm to do not ride the noise caused by the contact condition of the electrode.

We test four pattern motions, breathing, swallowing of water, deep breathing, vocalization. Four of the operation is a series of operations. Normal breathing, swallowing of water, take a deep breath performs each operation three times, utterance underlying vowel (/a/, /i/, /u/, /e/, /o/) uttering. We try each five times of this operations in three subjects. We perform the wavelet transform to analyze sEMG signals. Wavelet transform and k-NN method described in the next chapter. Here, it will show the overall flow of the experiment in Fig. 1.

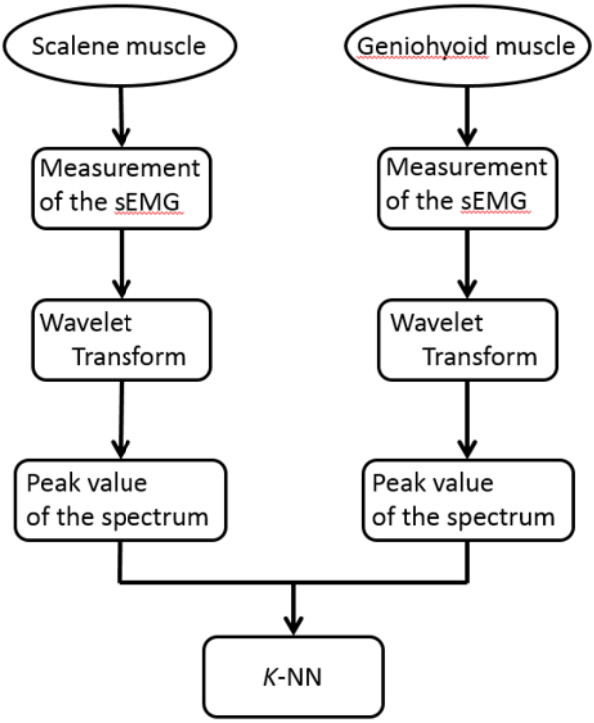


Fig. 1 Flow experiments

3 Analytical Method

3.1 Wavelet Transform

The wavelet transform is performed to extract features of the frequencies from the sEMG. Wavelet transform is extract the characteristics of the original signal by scaling and translating the underlying waveform (mother wavelet). The signal after wavelet transform is shown in Fig. 2 and Fig. 3. As a method, is performed averaging of the spectrum be-tween 100 msec of each operation after the Wavelet conversion, the peak of the spectrum of each operation can see the difference in turn many frequency. Fig. 4 shows that the spectrum of averaging between 100 msec of each operation. The vertical axis in Fig. 4 is a spectrum. From Fig. 4, the highest frequency peak value of the spectrum of frequencies is 125Hz. For this reason, we use the signal wave of 125Hz for the analysis.

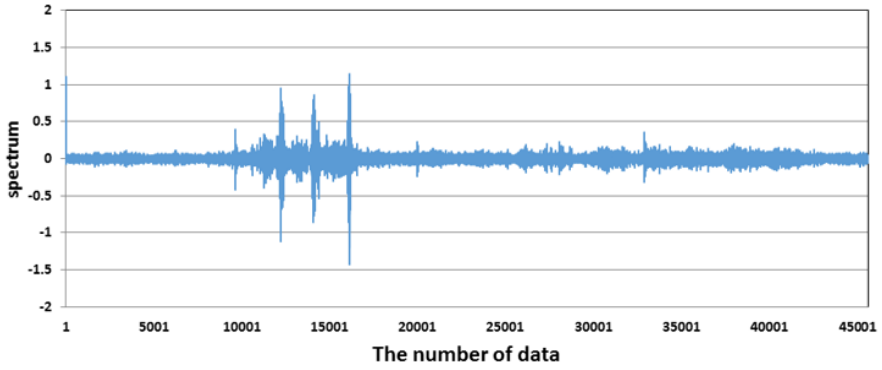


Fig. 2 Musculus geniohyoideus

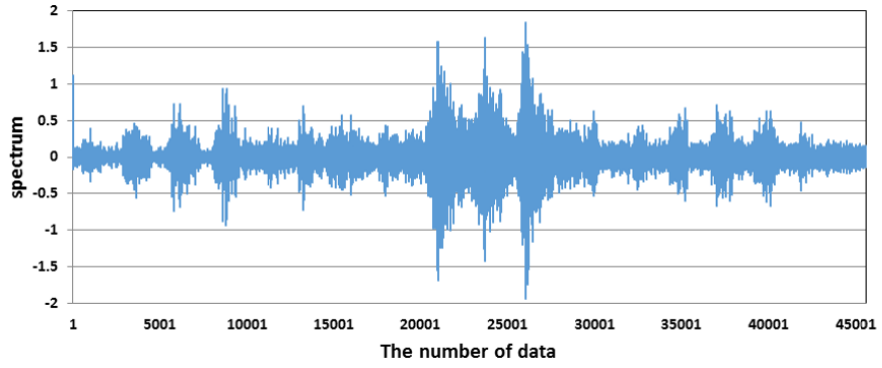


Fig. 3 Scalene muscle

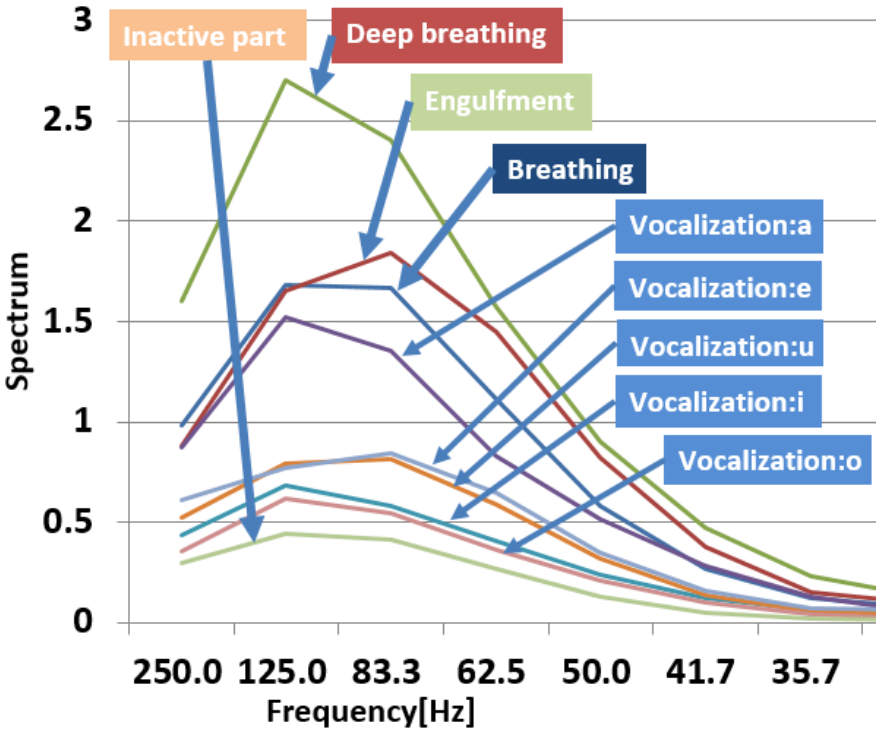


Fig. 4 The average spectrum between 100 msec

3.2 k-Nearest Neighbor Algorithm

K-NN method is a method for determining whether a majority vote using the k number of neighboring points from the data of its which unknown data operation when placed in unknown data or the group as a base set. In this paper, it is determined whether the data of each operation is in the set of proper operation to make the data that is first made mother set. Here we show the formula for the identification rate of the operation below.

$$\text{Classification rate} = \text{Success data} / \text{The total number of data} \quad (1)$$

Number of success divided by the total number of data is Equation (1). A plot of each data from wavelet transform are shown in Fig. 5, Fig. 6 and Fig. 7. The vertical axis represents the value from geniohyoid. The horizontal axis is the value from scalene muscle.

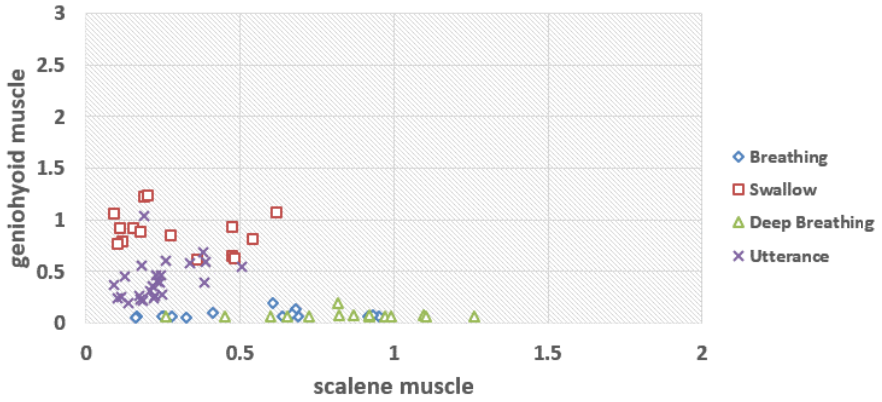


Fig. 5 Correlation diagram of subject A

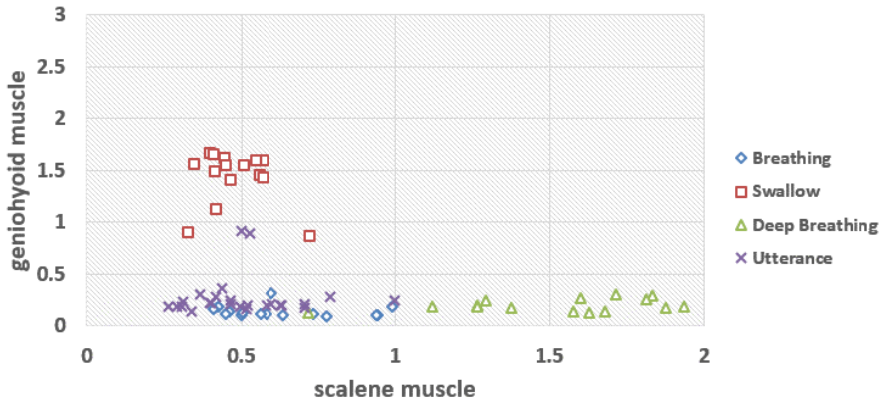


Fig. 6 Correlation diagram of subject B

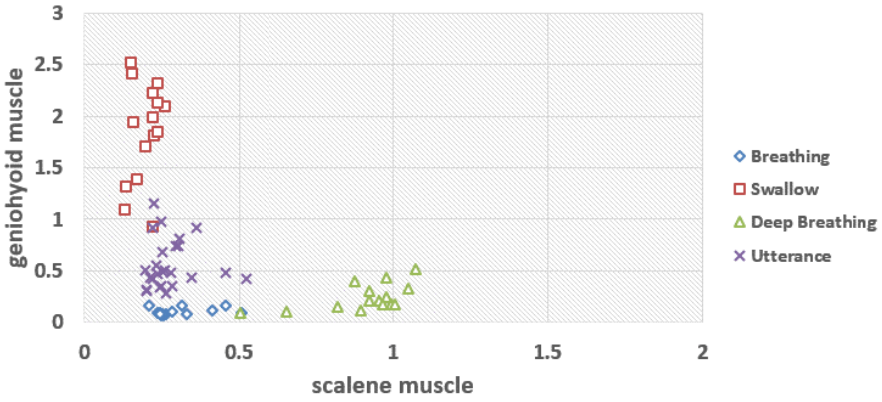


Fig. 7 Correlation diagram of subject C

4 Experimental Results

Experimental results show Fig. 8 and Fig. 9 and Fig. 10. The vertical axis is pattern classification rate. The horizontal axis is the number of k of the k-NN method.

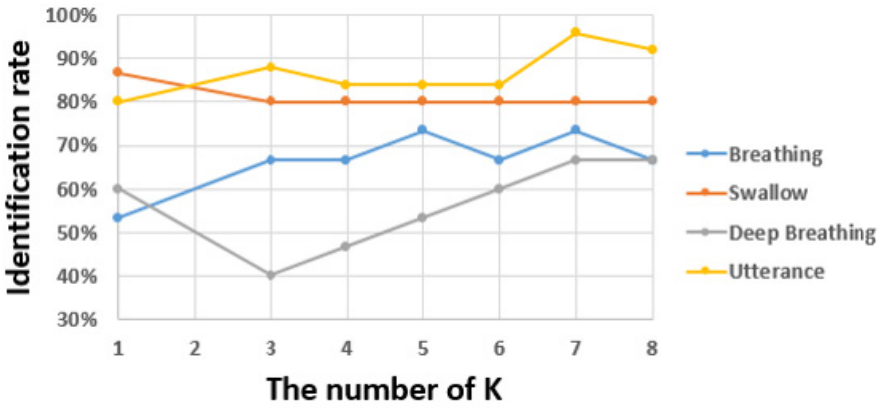


Fig. 8 Experimental results of subject A

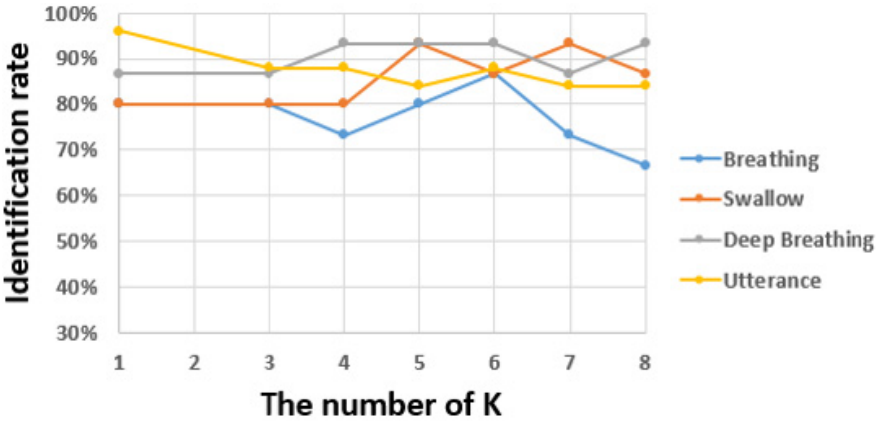


Fig. 9 Experimental results of subject B

In addition, we considered that the deep breath and the breathing are one breath pattern. The results are shown in Fig. 11, Fig. 12 and Fig. 13. The vertical axis is the classification rate. The horizontal axis is the number of k. From these results, classification rate became good results when k is 6, focusing on breathing.

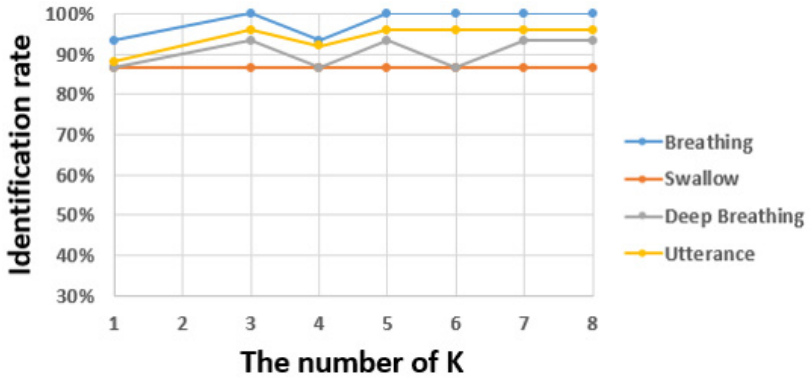


Fig. 10 Experimental results of subject B

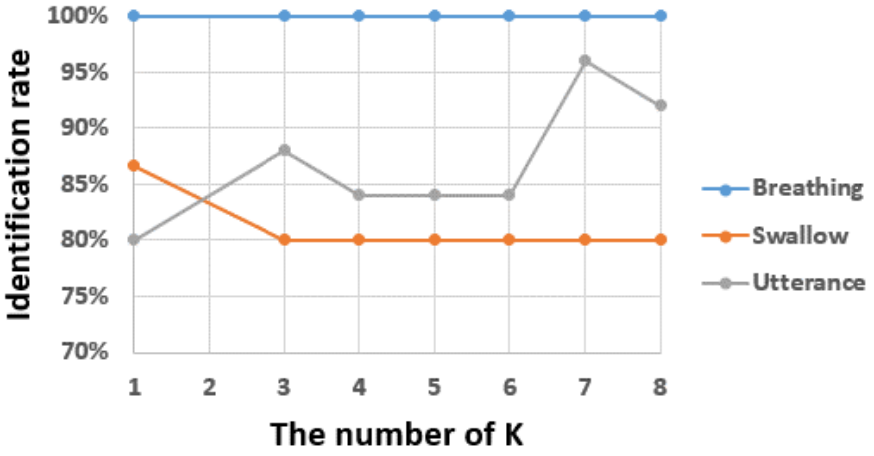


Fig. 11 3 pattern classification results of subject A

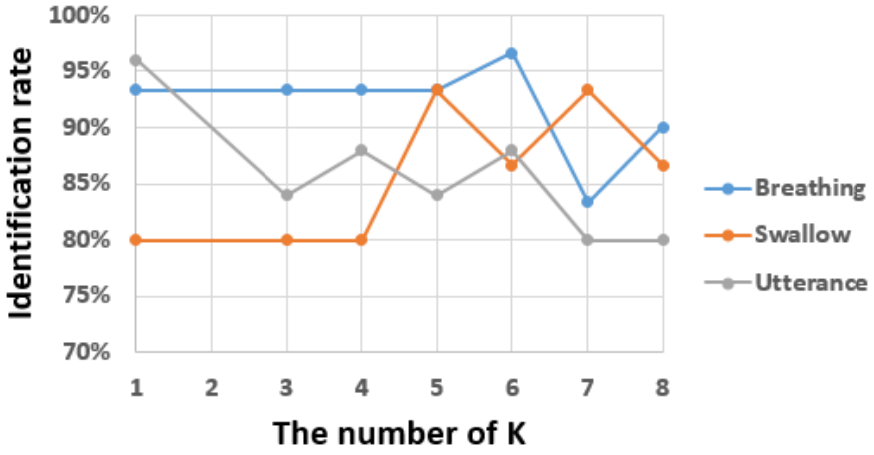


Fig. 12 3 pattern classification results of subject B

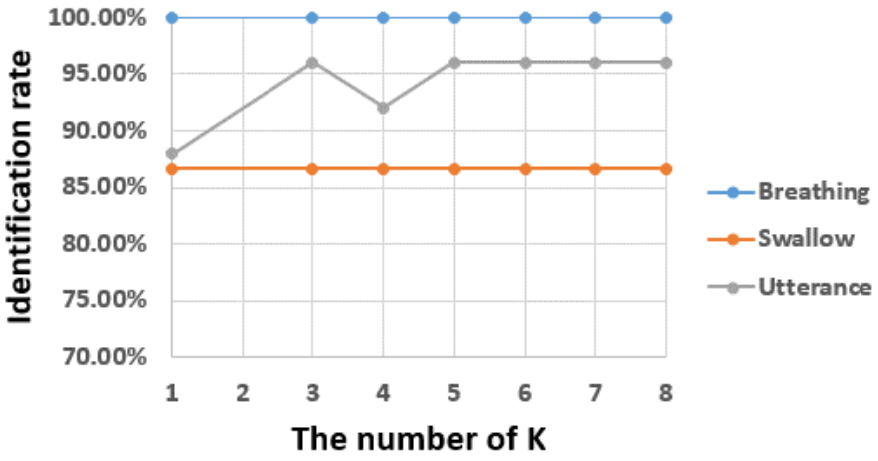


Fig. 13 3 pattern classification results of subject C

5 Conclusion

In this paper, we examined the method of pattern classification using sEMG in order to perform the diagnosis and treatment of sleep apnea syndrome. From experimental results, our analysis method had about 97% of the breathing pattern classification rate. In future works, it is necessary to increase the number of subjects.

References

1. Kizuka, T., Masuda, T., Kiryu, T., Sadoyama, T.: Biomechanism Library -Practical Usage of Surface Electromyogram-
2. Okitsu, T., Arita, M., Sonoda, S., Ota, T., Hotta, F., Honda, T., Chino, N.: The Surface Electromyography on Suprahyoid Muscles during Swallowing

High Power Wireless Power Transfer Driven by Square Wave Inputs

Kazuya Yamaguchi, Takuya Hirata and Ichijo Hodaka

Abstract A power source used in wireless power transfer generates an AC wave to transfer electric power to a load which is not connected electrically but connected electromagnetically to the power source. In this paper, the power and efficiency are compared when a sinusoidal and a square waves which are typical AC waves are applied as power source voltage outputs. The condition under which the two waves respectively transfer equivalent power with electric wires is examined to figure out the effect of different waves. Then the power and efficiency are calculated by a mathematical approach with practical values of elements on various situations. Finally, this paper explores how input waves should be chosen for ideal wireless power transfer.

Keywords Wireless power transfer · Resonant phenomenon · State space representation

1 Introduction

The technique of transmitting power enough to drive electronic devices such as LEDs without electric wires is called wireless power transfer(WPT) and has been noticed in recent years. Possibility of practical use of WPT is recognized with the successful experiment where the gap of wireless transmission was 0.6 meters reported in [1]. After that, the studies have reported WPT by various viewpoints. In [2], efficiency of transmission is expressed in terms of Q factors of circuits, and in [3] it is represented by Q factors with series and parallel circuits of transmitting and

K. Yamaguchi(✉) · T. Hirata · I. Hodaka
Interdisciplinary Graduate School of Agriculture and Engineering, University of Miyazaki, Miyazaki, Japan

K. Yamaguchi · T. Hirata · I. Hodaka
Department of Environmental Robotics, University of Miyazaki, Miyazaki, Japan
e-mail: {nc13001,nc14001}@student.miyazaki-u.ac.jp, hijhodaka@cc.miyazaki-u.ac.jp

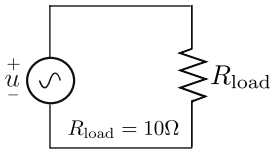
T.T. Zin et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 388,
DOI: 10.1007/978-3-319-23207-2_34

receiving. [4] derives the expressions of efficiency and the optimal load resistance with relay circuit between the transmitting side and the receiving side, and [5] shows that the use of a relay circuit serves a high Q value. In addition, [6] analyzes a WPT circuit by an equivalent circuit, and [7] investigates the electromagnetic field with resonant phenomena related to WPT analysis.

Most of researches on WPT uses sinusoidal waves as voltage at the transmitting power source. Some researches use square waves instead of sinusoidal waves, since generating square waves is much easier and more cost-effective. However, little is known about WPT with square wave inputs. [8] applies a square wave as the power source voltage, and observes WPT by approximating the square wave as a sinusoidal wave. The approximation is possible if the circuit has a single resonant frequency and the power source voltage can be driven at the resonant frequency of the circuit; however, it should be verified under the other situations. This paper argues the ability of WPT circuits with different types of input; we figure out power and efficiency of WPT circuits with sinusoidal and square wave inputs.

2 Preliminaries

In this paper, we are interested in how a sinusoidal-wave input and a square-wave input respectively affect the outputs of WPT circuits, and what difference about performance of WPT system there is. The amplitude of inputs are adjusted in such a way that when they are used in the circuit in Fig. 1 the load powers of those are equal. This situation makes the two wave inputs have ability to deliver the same power. In the rest of the paper, the power supply voltages which are defined as in Fig. 1 if they are not specified.



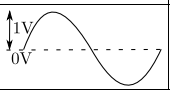
	sinusoadal	square
Power	50.0mW	50.0mW
Efficiency	1.00	1.00
Amplitude		

Fig. 1 Power and efficiency with different inputs

3 Basic Wireless Power Transfer Circuits

This section discusses choice of input wave for the WPT circuit in Fig. 2. Then we focus on particular situation where using square waves as input voltage surpasses pure sinusoidal waves.

3.1 Circuit Equation and Transfer Function

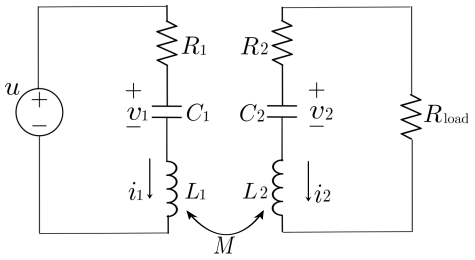
On the WPT circuit in Fig. 2, the left side is a transmitting circuit and the right side is a receiving circuit, and the power which is generated at power source is delivered wirelessly and then consumed at a load. The values R_1, R_2, C_1, and C_2 are parasitic factors of the circuit and inductances, R_{load} is a load resistance, L_1 and L_2 are self inductances, and M is a mutual inductance. For the circuit, the state space equation is derived as (1) and the transfer function from the voltage of power source u to the current of receiving side i_2 (2) can be obtained (but omitted here)[9].

$$\begin{aligned} &\dot{x} = Ax + Bu,\ x = \begin{bmatrix} v_1 & v_2 & i_1 & i_2 \end{bmatrix}^{\mathrm{T}} \\ &A = \frac{1}{\Delta}\begin{bmatrix} 0 & 0 & \frac{\Delta}{C_1} & 0 \\ 0 & 0 & 0 & \frac{\Delta}{C_2} \\ -L_2 & M & -R_1L_2 & R_3M \\ M & -L_1 & R_1M & -R_3L_1 \end{bmatrix},\ B = \frac{1}{\Delta}\begin{bmatrix} 0 \\ 0 \\ L_2 \\ -M \end{bmatrix} \\ &\Delta = L_1L_2 - M^2,\ R_3 = R_2 + R_{\text{load}} \end{aligned} \tag{1}$$

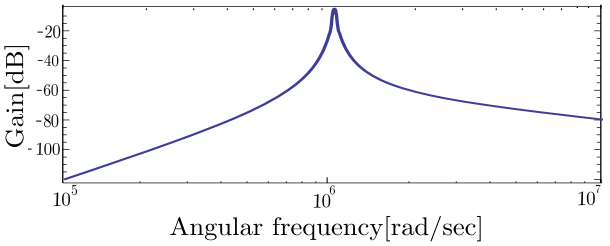
$$\begin{aligned} G(s) &= \frac{-MC_1C_2s^3}{D(s)} \\ D(s) &= \Delta C_1C_2s^4 + (R_1L_2 + R_3L_1)C_1C_2s^3 \\ &\quad + (L_1C_1 + L_2C_2 + R_1R_3C_1C_2)s^2 + (R_1C_1 + R_3C_2)s + 1 \end{aligned} \tag{2}$$

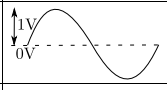
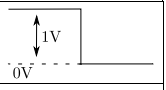
3.2 Sinusoidal Wave or Square Wave as Input

In this subsection, practical values of resistances, capacitances, self inductances, and mutual inductances are used. With these values, power and efficiency are calculated with a sinusoidal and square waves which have the same frequency based on the equation (1). The power and efficiency with inputs whose frequency are fixed at the resonant frequency of the circuit and at one third are shown in Fig. 2. If a sinusoidal wave at the resonant frequency is used, the output power is maximized because the current is proportional to the voltage at the resistive load. On the left table in Fig. 2, we can see that WPT driven by the sinusoidal wave is better than by the square wave at the resonant frequency. However if the frequency of the power supply is restricted to be less than the resonant frequency, it is not the case. On the right table in Fig. 2, the frequency is set to one third of the resonant frequency. In this case, both power and efficiency by the square wave input are higher than by the sinusoidal wave input. One reason is that square waves are represented by Fourier series which are composed of many sinusoidal waves with different frequencies. In [8], using the resonant frequency is suitable when they use square waves well as sinusoidal waves.



R_1	0.1Ω	L_1, L_2	100μH
R_2	1Ω	M	1μH
R_{load}	10Ω	C_1, C_2	10nF



	sinusoadal	square
P_1	2.62mW	1.06mW
P_2	1.13mW	0.460mW
η	0.433	0.433
Amplitude		
Angular Frequency	10^6rad/sec	10^6rad/sec

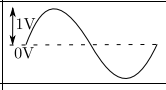
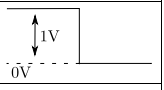
	sinusoadal	square
P_1	703nW	118mW
P_2	0.110nW	51.1mW
η	1.56×10^{-4}	0.433
Amplitude		
Angular Frequency	3.33×10^5rad/sec	3.33×10^5rad/sec

Fig. 2 The numerical examples

However, it is revealed here that the square waves skillfully utilizes the resonant phenomena in this case.

4 Use of a Relay Circuit to Transfer Power

In this section, the difference with a sinusoidal wave input and with a square wave input is examined by the WPT circuit with a relay of Fig. 3. On the WPT circuit of Fig. 3, the left side is a transmitting circuit, the middle side is a relay circuit, and the right side is a receiving circuit. It is the characteristic that Fig. 3 has a few resonant frequencies for the circuit of Fig. 2 which has a single resonant frequency. The state space model of the circuit is shown as the equation (3)[10], and the Bode diagram is shown in Fig. 3. The Bode plot has three peaks at three resonant frequencies since there are three inductances and capacitances respectively. In this example, high power can be obtained if a square wave input whose frequency is adjusted to the lowest resonant frequency drives the circuit.

$$\dot{x} = Ax + Bu,\ x = \left[\, v_1\ v_2\ v_3\ i_1\ i_2\ i_3 \,\right]^{\mathrm{T}} \tag{3}$$

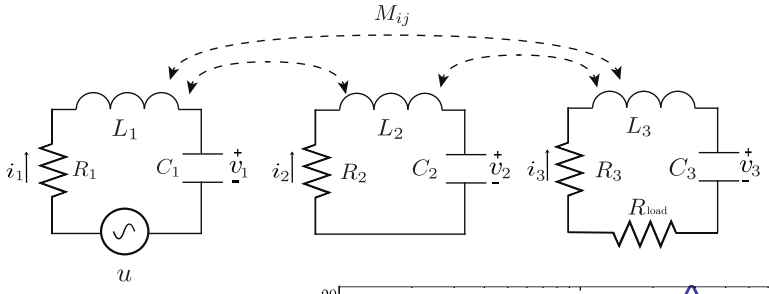
$$A = \frac{1}{\Delta}\begin{bmatrix} 0 & 0 & 0 & \frac{\Delta}{C_1} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{\Delta}{C_2} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{\Delta}{C_3} \\ -\Lambda_{23} & \Gamma_3 & \Gamma_2 & -\Lambda_{23}R_1 & \Gamma_3 R_2 & \Gamma_2 R_4 \\ \Gamma_3 & -\Lambda_{31} & \Gamma_1 & \Gamma_3 R_1 & -\Lambda_{31}R_2 & \Gamma_1 R_4 \\ \Gamma_2 & \Gamma_1 & -\Lambda_{12} & \Gamma_2 R_1 & \Gamma_1 R_2 & -\Lambda_{12}R_4 \end{bmatrix},\ B = \frac{1}{\Delta}\begin{bmatrix} 0 \\ 0 \\ 0 \\ \Lambda_{23} \\ -\Gamma_3 \\ -\Gamma_2 \end{bmatrix}$$

$$\Delta = L_1L_2L_3 - L_1M_{23}^2 - L_2M_{31}^2 - L_3M_{12}^2 + 2M_{12}M_{23}M_{31}$$

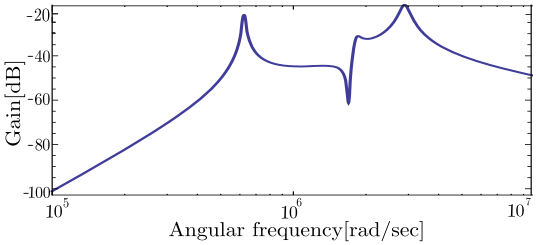
$$\Lambda_{12} = L_1L_2 - M_{12}^2,\ \Lambda_{23} = L_2L_3 - M_{23}^2,\ \Lambda_{31} = L_3L_1 - M_{31}^2$$

$$\Gamma_1 = L_1M_{23} - M_{12}M_{31},\ \Gamma_2 = L_2M_{31} - M_{12}M_{23},\ \Gamma_3 = L_3M_{12} - M_{31}M_{23}$$

$$R_4 = R_3 + R_{\mathrm{load}}$$



R_1, R_2, R_3	0.1Ω	M_{12}	80μH
R_{load}	10Ω	M_{23}	70μH
L_1, L_2, L_3	100μH	M_{31}	86μH
C_1, C_2, C_3	10nF		



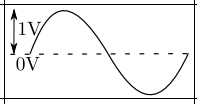
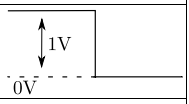
	sinusoadal	square
P_1	33.3mW	21.8mW
P_2	20.6mW	21.2mW
η	0.618	0.971
Amplitude		
Angular Frequency	6.24×10^5rad/sec	6.24×10^5rad/sec

Fig. 3 Relay circuit

5 Conclusion

In this paper, we have discussed how the type of inputs affect the power and efficiency of a WPT circuit on the standard WPT circuit and with the relay circuit. In general, high output power can be obtained by using a sinusoidal wave and adjusting its frequency if a WPT circuit has a simple resonant frequency. In contrast, it has been revealed that a square wave should be adopted rather than a sinusoidal wave when the available frequency of power source is confined or the WPT circuit has plural resonant frequencies.

References

1. Kurs, A., Karalis, A., Moffatt, R., Joannopoulos, J.D., Fisher, P., Soljačić, M.: Wireless Power Transfer via Strongly Coupled Magnetic Resonances. Science **317**, 83–86 (2007)
2. Tucker, C.A., Warwick, K., Holderbaum, W.: A contribution to the wireless transmission of power. Electrical Power and Energy Systems **47**, 235–242 (2013)
3. Vandevoorde, G., Puers, R.: Wireless energy transfer for stand-alone systems: a comparison between low and high power applicability. Sensors and Actuators A **92**, 305–311 (2001)
4. Takura, T., Misawa, T., Sato, F., Sato, T., Matsuki, H.: Maximum Transmission Efficiency of LC-Booster Using Pick-up Coil with Capacitance. Journal of the Magnetics Society of Japan **37**, 102–106 (2013)
5. Hoang, H., Bien, F.: Maximizing efficiency of electromagnetic resonance wireless power transmission systems with adaptive circuits. In: Wireless Power Transfer - Principles and Engineering Explorations, pp. 207–226 (2012)
6. Imura, T., Hori, Y.: Maximizing Air Gap and Efficiency of Magnetic Resonant Coupling for Wireless Power Transfer Using Equivalent Circuit and Neumann Formula. IEEE Transactions Industrial Electronics **58**, 4746–4752 (2011)
7. Tan, L., Huang, X., Huang, H., Zou, Y., Li, H.: Transfer efficiency optimal control of magnetic resonance coupled. Science China Technological Sciences **54**, 1428–1434 (2011)
8. de Boeij, J., Lomonova, E., Duarte, J.L., Vandenput, A.J.A.: Contactless power supply for moving sensors and actuators in high-precision mechatronic systems with long-stroke power transfer capability in x-y plane. Sensors and Actuators A **148**, 319–328 (2008)
9. Yamaguchi, K., Hirata, T., Yamamoto, Y., Hodaka, I.: Resonance and efficiency in wireless power transfer system. WSEAS Transactions on Circuits and Systems **13**, 218–223 (2014)
10. Hirata, T., Yamamoto, Y., Yamaguchi, K., Setiawan, E., Hodaka, I.: On circuit topologies of wireless power transmission with relay coils. In: Proceedings of The 3rd International Conference on Computer Engineering & Mathematical Sciences, pp. 200–202 (2014)

Analyzing Tagging Accuracy of Part-of-Speech Taggers

Nyein Pyae Pyae Khin and Than Nwe Aung

Abstract Automated part-of-speech (POS) tagging has been a very active research area for many years and is the foundation of natural language processing systems. Natural Language Toolkit (NLTK) library in the Python environment provides the necessary tools for tagging, but doesn't actually tell us what methods work the best. Therefore, this work analyzes the performance of part-of-speech taggers, namely the NLTK Default tagger, Regex tagger and N-gram taggers (Unigram, Bigram and Trigram) on a particular corpus. The corpora we have used for the analysis are; Brown, Penn Treebank and CoNLL2000. We have applied all taggers to these three corpora, resultantly we have shown that whereas Unigram tagger does the best tagging in all corpora, the combination of taggers does better if it is correctly ordered.

Keywords POS taggers · Brown corpus · Penn Treebank Corpus · CoNLL2000 corpus

1 Introduction

Part-Of-Speech (POS) tagging is the process of identifying nouns, verbs, adjectives, pronouns, conjunction and other parts of speech in context. POS tagging can be used for Linguistic-text pre-processing before semantic analysis. Research on part-of-speech tagging has been closely tied to corpus linguistics. The tag sets, the collection of tags used for a particular task and the terminology of tagging are defined by different projects such as Brown Corpus Tag-Set, Penn Treebank Tag-Set. In Python programming language there is an infrastructure about natural language processing, which is called Natural Language Toolkit (NLTK). NLTK includes extensive software, data, and documentation, all freely downloadable from *http://www.nltk.org/*. Distributions are provided for Windows, Macintosh,

N.P.P. Khin(✉) · T.N. Aung(✉)
University of Computer Studies, Mandalay, Myanmar
e-mail: {nyeinpyaepyaekhin,mdytna}@gmail.com

T.T. Zin et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 388,
DOI: 10.1007/978-3-319-23207-2_35

and Unix platforms [1]. The Natural Language Toolkit (NLTK) is a Python package for natural language processing. NLTK requires Python 2.6, 2.7, or 3.2+.

In the Python development environment, there are several tagged corpora available for installation. The available corpus names are listed by using *nltk.download( )* method [1]. However, not all of the corpora listed by this method are tagged. In this paper, we will be dealing with tagged corpora, *i.e*. which includes part-of-speech annotations. The tagged corpora installed for this work in the Python NLTK library are as follows; Brown Corpus[9], Penn Treebank Corpus [6]and CoNLL2000 Corpus[10]. The aim of this paper is to compare and evaluate POS taggers, namely NLTK Default tagger, Regex tagger, N-gram taggers (Unigram, Bigram and Trigram) and tagger combination on the selected corpus.

In this study, we have more deeply examined the different corpora and taggers available in NLTK. We have first listed POS taggers used for analyzing tagging accuracy. The Python source code for analyzing the CoNLL2000 corpus is written in method and analysis section to show the exact way of how we made the analysis. While writing the code for analysis, we have inspired from the code samples used in [1]. Finally, all corpora are analyzed by using the same tagging methods and the accuracies are listed in Table 1.

The paper proceeds as follows: Section 2 describes POS taggers and gives an overview of each tagger. Section 3 details measuring of tagging accuracy. Section 4 reports experimental result. Finally, in Section 5, concludes the paper.

2 POS Taggers

There are three main POS taggers that we will use:

1. ***Default Tagger***: This tagger tags every word with a default tag. For example, a very good baseline for English POS-tagging is to just tag every word as a noun.

2. ***Regex Tagger***: This tagger uses human-defined patterns and tags according to that pattern array.

3. ***N-gramTaggers***: N-grams over any given sequence can be informally defined as overlapping subsequences each of length N. As an example, the sentence “My name is Nyein Pyae” will yield the following n-grams for various values of N:
   - N = 1 (1-grams or Unigrams): My, name, is, Nyein, Pyae
   - N = 2 (2-grams or Bigrams): My name, name is, is Nyein, Nyein Pyae
   - N = 3 (3-grams or Trigrams): My name is, name is Nyein, is Nyein Pyae

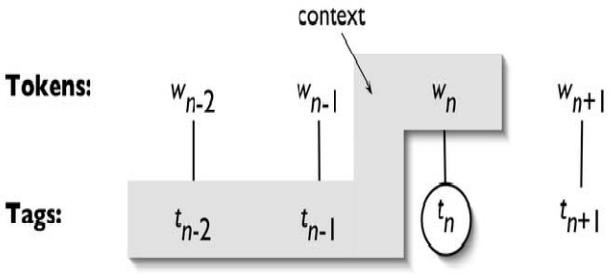


Fig. 1 N-gram tagger context [1]

Fig. 1. shows the basic idea behind this strategy. Instead of just looking at the word being tagged, we also look at the POS tags of the previous *n* words. Therefore, using n-grams allows us to be able to take context into consideration when performing POS-tagging. The important thing to realize is that when using an N-gram Tagger, we need to train it on some sentences for which we already know the POS tags. This is needed because an N-gram Tagger needs to count and build tables of how many times a particular word is tagged as a verb (when N=1) or how many times a particular word preceded by a noun is tagged as a verb (when N=2) and so on.

3 Measuring Accuracy of Tagging

A POS tagger attempts to assign the correct POS tag or lexical category to all words of a given text, usually by relying on the assumption that a word can be assigned a single POS tag by looking at the POS tags of the neighbouring words. The source code used for corpus exploration and the POS tagger applications are explained by using the CoNLL2000 corpus as follows:

```
Python 3.4.3 Shell
File  Edit  Shell  Debug  Options  Window  Help
Python 3.4.3 (v3.4.3:9b73f1c3e601, Feb 24 2015, 22:43:06) [MSC v.1600 32 bit (Intel)] on win32
Type "copyright", "credits" or "license()" for more information.
>>> import nltk
>>> from nltk.corpus import conll2000
>>> tagged=conll2000.tagged_words()
>>> tagged
[('Confidence', 'NN'), ('in', 'IN'), ('the', 'DT'), ...]
>>> con_tagged_sents=conll2000.tagged_sents()
>>> con_tagged_sents
[[('Confidence', 'NN'), ('in', 'IN'), ('the', 'DT'), ('pound', 'NN'), ('is', 'VBZ'), ('widely', 'RB'), ('expected', 'VBN'), ('to', 'TO'), ('take', 'VB'), ('another', '
DT'), ('sharp', 'JJ'), ('dive', 'NN'), ('if', 'IN'), ('trade', 'NN'), ('figures', 'NNS'), ('for', 'IN'), ('September', 'NNP'), (',', ','), ('due', 'JJ'), ('for', 'IN')
, ('release', 'NN'), ('tomorrow', 'NN'), (',', ','), ('fail', 'VB'), ('to', 'TO'), ('show', 'VB'), ('a', 'DT'), ('substantial', 'JJ'), ('improvement', 'NN'), ('from',
'IN'), ('July', 'NNP'), ('and', 'CC'), ('August', 'NNP'), ("'s", 'POS'), ('near-record', 'JJ'), ('deficits', 'NNS'), ('.', '.')], [('Chancellor', 'NNP'), ('of', 'IN'),
('the', 'DT'), ('Exchequer', 'NNP'), ('Nigel', 'NNP'), ('Lawson', 'NNP'), ("'s", 'POS'), ('restated', 'VBN'), ('commitment', 'NN'), ('to', 'TO'), ('a', 'DT'), ('firm',
'NN'), ('monetary', 'JJ'), ('policy', 'NN'), ('has', 'VBZ'), ('helped', 'VBN'), ('to', 'TO'), ('prevent', 'VB'), ('a', 'DT'), ('freefall', 'NN'), ('in', 'IN'), ('sterl
ing', 'NN'), ('over', 'IN'), ('the', 'DT'), ('past', 'JJ'), ('week', 'NN'), ('.', '.')], ...]
>>> con_sents=conll2000.sents()
>>> con_sents
[['Confidence', 'in', 'the', 'pound', 'is', 'widely', 'expected', 'to', 'take', 'another', 'sharp', 'dive', 'if', 'trade', 'figures', 'for', 'September', ',', 'due', '
for', 'release', 'tomorrow', ',', 'fail', 'to', 'show', 'a', 'substantial', 'improvement', 'from', 'July', 'and', 'August', "'s", 'near-record', 'deficits', '.'], ['Ch
ancellor', 'of', 'the', 'Exchequer', 'Nigel', 'Lawson', "'s", 'restated', 'commitment', 'to', 'a', 'firm', 'monetary', 'policy', 'has', 'helped', 'to', 'prevent', 'a',
'freefall', 'in', 'sterling', 'over', 'the', 'past', 'week', '.'], ...]
>>>
```

We have found that not all of the corpora installable are tagged. Then we filtered the tagged corpora as if the tagged words( ) method is reachable in a corpus, we assumed it as a tagged corpus. The assumption is made according to the NLTK web site, which is written as tagged words( ) method is only supported by corpora that include part-of-speech annotations [12]. As a result of this filtering, we have found that Brown Corpus, Penn Treebank Corpus and CoNLL2000 Corpus are available with their tagged sentences. After all, we have applied Default Tagger, Regex Tagger, Bigram Tagger, Unigram Tagger, Trigram Tagger and Combination of taggers to these three corpora. The Python source code for analyzing the CoNLL2000 corpus is shown in the following sessions.

3.1 Default Tagger

Firstly, in order to evaluate the taggers on the CoNLL2000 corpus, "Default tagger" is analyzed. The default tagger tags each token with the most common tag. In order to get the best result, we tag each word with the most likely tag (i.e. NN tag) [1]. In the program code below shows that 14.20% of the tags are nouns for the CoNLL2000 corpus.

```
>>> defaultTagger=nltk.DefaultTagger('NN')
>>> evalR=defaultTagger.evaluate(conll2000.tagged_sents())
>>> print ("Accuracy of default tagger is: %4.2f %%" % (100.0 * evalR))
Accuracy of default tagger is: 14.20 %
>>>
```

We can create a tagger, called *defaultTagger*, this tags everything as NN. Default taggers assign their tag to every single word, even words that have never been encountered before. As it happens, most new words are nouns. Thus, default taggers help to improve the robustness of a language processing system.

3.2 Regular Expression Tagger

The regular expression (Regex) tagger assigns tags to tokens on the basis of matching human defined patterns. For instance, we might guess that any word ending in '*ed*' is the past participle of a verb, and any word ending with '*'s*' is a possessive noun [1]. We can express these as a list of regular expressions:

```
>>> patterns=[
        (r'.*ing$', 'VBG'),                 # gerunds
        (r'.*ed$', 'VBD'),                  # simple past
        (r'.*es$', 'VBZ'),                  # 3rd singular present
        (r'.*ould$', 'MD'),                 # modals
        (r'.*\'s$', 'NN$'),                 # possessive nouns
        (r'.*s$', 'NNS'),                   # plural nouns
        (r'^-?[0-9]+(.[0-9]+)?$', 'CD'),    # cardinal numbers
        (r'.*', 'NN')                       # nouns (default)
        ]
>>>
```

Regex tagger does a better job of handling "unseen" words than the 'NN' default tagger. It is More sophisticated than the 'NN' default tagger. This tagger by itself is limited to very common language properties; therefore it is able to tag only the 22.11% of the whole corpus correctly.

```
>>> regexTagger = nltk.RegexpTagger(patterns)
>>> evalResult = regexTagger.evaluate(conll2000.tagged_sents())
>>> print ("Accuracy of Regex tagger is: %4.2f %%" % (100.0 * evalResult))
Accuracy of Regex tagger is: 22.11 %
>>>
```

3.3 N-Gram Taggers

In general, an n-gram tagger makes a decision for a given word, one at a time, in a single direction. We can analyzed each of the 3 N-gram Taggers: Unigram Tagger, Bigram Tagger, and Trigram Tagger as shown follows:

3.3.1 Unigram Tagger

The unigram tagger implements a simple statistical tagging algorithm: for each token, it assigns the tag that is most likely for that token's type. Before a unigram tagger can be used to tag data, it must be trained on a training corpus. It uses this corpus to determine which tags are most common for each word. Unigram Tagger is applied to the CoNLL2000 corpus with the Python code below and it is able to tag 89.47% of the whole corpus correctly.

```
>>> corpusSize = int(len(conll2000.tagged_sents()) * 0.8)
>>> trainingSents = conll2000.tagged_sents()[:corpusSize]
>>> testSents = conll2000.tagged_sents()[corpusSize:]
>>> unigram_tagger = nltk.UnigramTagger(trainingSents)
>>> evalResult = unigram_tagger.evaluate (testSents)
>>> print ("Accuracy of Unigram tagger is: %4.2f %%" % (100.0 * evalResult))
Accuracy of Unigram tagger is: 89.47 %
>>>
```

3.3.2 Bigram Tagger

The bigram tagger is applied similar to the unigram tagger, the same training and test sets are used. The accuracy of the tagger (20.31%) decreases dramatically according to the unigram tagger. This decrease is explained in [1] as follows; Notice that the bigram tagger manages to tag every word in a sentence it saw during training, but does badly on an unseen sentence.

```
>>> bigramTagger = nltk.BigramTagger(trainingSents)
>>> evalResult = bigramTagger.evaluate(testSents)
>>> print ("Accuracy of Bigram tagger is: %4.2f %%" % (100.0 * evalResult))
Accuracy of Bigram tagger is: 20.31 %
>>>
```

3.3.3 Trigram Tagger

We applied the trigram tagger to CoNLL2000 corpus. The trigram tagger will be using the part-of-speech tag of the previous two tokens, which will normally be the last word of the previous sentence and the sentence-ending punctuation [1]. However, the lexical category that closed the previous sentence has no bearing on the one that begins the next sentence. As it is seen in the following code, the same training and test sets are applied. The accuracy of the trigram tagger (10.87%) is also decrease than the bigram tagger for CoNLL2000 corpus.

```
>>> TrigramTagger = nltk.TrigramTagger(trainingSents)
>>> evalResult = TrigramTagger.evaluate(testSents)
>>> print ("Accuracy of Trigram tagger is: %4.2f %%" % (100.0 * evalResult))
Accuracy of Trigram tagger is: 10.87 %
>>>
```

3.4 Combination of Taggers

It is possible to combine taggers such that if the primary tagger was unable to assign the tag to a particular word, it backs off to the second tagger for the prediction [2]. This is known as *Backoff*. Most NLTK taggers permit a backoff tagger to be specified. For each token to be tagged, the backoff tagger consults each subtagger. Thus, we applied all the taggers by combining them in an order as follows:

```
>>> defaultTagger = nltk.DefaultTagger('NN')
>>> regexpTagger = nltk.RegexpTagger(patterns, backoff = defaultTagger)
>>> unigramTagger = nltk.UnigramTagger(trainingSents, backoff = regexpTagger)
>>> bigramTagger = nltk.BigramTagger(trainingSents, backoff = unigramTagger)
>>> trigramTagger = nltk.TrigramTagger(trainingSents, backoff = bigramTagger)
>>> evalResult = trigramTagger.evaluate(testSents)
>>> print ("Accuracy of combination of taggers is: %4.2f %%" % (100.0 * evalResult))
Accuracy of combination of taggers is: 92.54 %
>>>
```

Note that the backoff sequence is in reverse order, so for the trigram tagger backs off to the bigram tagger, which backs off to the unigram tagger and so on**.** The accuracy of combination of taggers is 92.54%. Consequently, the result becomes better than all of the above taggers used.

4 Experimental Result

To test the accuracy of a part-of-speech tagger, we can compare it to the test sentences. The accuracy represents the percentage of words the taggers have tagged correctly. Table 1 displays the total amount of tokens in each corpora used for this experiment.

Table 1 Tagging accuracy of NLTK.

Corpus Name	Default	Regex	Unigram	Bigram	Trigram	Comb. Of Taggers
CoNLL2000	14.20 %	22.11 %	89.47 %	20.31 %	10.87 %	92.54 %
Brown	13.13 %	28.33 %	87.71 %	33.90 %	19.21 %	91.29 %
Treebank	13.08 %	20.53 %	86.35 %	11.34 %	6.71 %	89.62 %

As seen in Table 1, Unigram Tagger performs better than the other taggers when applied alone on all three corpora. But the combinations of taggers' results are even better. Trigram and bigram taggers give lower results in accuracy. The experiments conducted show that the bigram tagger performs better than the trigram tagger in all cases. When we look at the results vertically for the corpora, tagging accuracy of the Brown corpus is better than the other ones with the bigram and trigram taggers. Whereas, n-gram taggers in general give more accurate results than the others applied, bigram and trigram taggers need more data to train in order to give better results for the Treebank and CoNLL2000 Corpora.

5 Conclusion

In this work, part-of-speech tagging is analyzed by improving the usage of different techniques and different corpora. The differences between each corpus are analyzed according to their tagging accuracies. The most important component of part-of-speech tagging is using the correct training data. A tagger trained on the CoNLL2000 corpus will be accurate for the treebank corpus, and vice versa, because CoNLL2000 and Treebank are quite similar. So make sure you choose your training data carefully. The tagger works by comparing the text given to it to a corpus of pre-tagged text. For the future work, some more improvements will be made for the regular expression tagger by using more complex patterns.

References

1. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python. OReilly Media, USA (2009)
2. Boehm, I.: Unigram Backoff vs. TnT Evaluating Part of Speech Taggers, Introduction to Computational Linguistics, Austria

3. Smedt, T.D., Marfia, F., Matteucci, M., Daelemans, W.: Using Wiktionary to Build an Italian, CLiPS Computational Linguistics Research Group. University of Antwerp
4. Sheikh, Z.M.A.W.: A Trigram Part-of-Speech Tagger for the Apertium Free/Open Source Machine Translation Platform, Computer Science and Engineering. National Institute of Technology Allahabad-211004, India
5. Hagerman, C.: Evaluating the Performance of Automated Part-of-Speech Taggers on an L2 Corpus. Osaka Jogakuin College
6. Marcus, M.P., Santorini, B., Marcinkiewicz, M.A.: Building a Large Annotated Corpus of English: The Penn Treebank. Computational Linguistics **19**, 313–330 (1993)
7. Part-Of-Speech tagging with NLTK. https://streamhacker.wordpress.com/tag/tagging/
8. NLTK 3.0 Documentation. http://www.nltk.org/
9. Brown Corpus Manual. http://icame.uib.no/brown/bcm.html
10. NLTK Default Tagger Performance on CoNLL2000. http://streamhacker.com/2011/01/25/nltk-default-tagger-conll2000-tag-coverage/
11. Processing Corpora with Python and the Natural Language Toolkit. http://www.freecode.com/articles/processing-corpora-with-python-and-the-natural-language-toolkit
12. Corpus Readers-Tagged Corpora. http://www.nltk.org/howto/corpus.html#tagged-corpora

Detection of Airway Obstruction from Frequency Distribution Feature of Lung Sounds with Small Power of Abnormal Sounds

Tomoki Nakano and Shigeyoshi Nakajima

Abstract We propose a new method to detect airway obstruction from a lung sound record with power of which abnormal sounds is much smaller than power of normal sound s. One of traditional methods to detect airway obstruction is FEV1% (forced expiratory volume 1 sec percentage) using a spirometry. But it bothers a patient too much. Some methods were proposed recently to detect abnormal sounds because an airway obstruction sometimes makes abnormal sounds such as wheeze or rhonchi or else. But it is not available for cases with small power of abnormal sounds. The correlation coefficient between our proposed value and FEV1% was -.592. And the AUC value of the proposed method with 70% threshold of FEV1% was 0.833. The proposed method could detect airway obstruction with sensitivity=0.8 and specificity = 0.78 *FEV*1%.

Keywords Bronchial asthma · Airway constriction · *FEV*1% · Fourier transform · Wheeze · Diagnosis

1 Introduction

1.1 Recent Works

There are many people annoyed by bronchial asthma [1-2]. The number of such people increases these days . An auscultation of lung sounds by a doctor is an

T. Nakano
Mitsubishi Electric Information Systems Corporation, Tokyo, Japan
e-mail: dokodemoikerukittoikeru@gmail.com

S. Nakajima(✉)
Grad School of Engneering, Osaka City University, Osaka, Japan
e-mail: nakajima@info.eng.osaka-cu.ac.jp

T.T. Zin et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 388,
DOI: 10.1007/978-3-319-23207-2_36

effective standard method for bronchial asthma. But an airway obstruction from bronchial asthma often occurs at the onset of sleep and often doesn't occur in front of a doctor in daytime . Then automatic detections of airway obstructions from lung sounds recorded in a home along 24 hours in a day seems very useful.

A wheeze is an abnormal noise often occurs during airway obstruction. Some lung sounds analyzer were proposed [3-11]. Most of them detect airway obstruction using rate of wheeze time per total time in a breath. Cases which rate are over a threshold become positive in detection of airway obstruction. But those methods are not available with cases without explicit wheeze for a doctor's ear. Wheeze is made from a vibration when air goes through a narrowed airway. But wheeze occurs randomly. There are some cases of airway obstruction without explicit wheeze. While wheeze often occurs in severe patients, this is not always the case in mild asthma. So conventional lung sound analyses are available for severe cases but not for mild cases.

There are traditional method s without computational analyses. One of them is *FEV*1% measurement. A simple spirometer is used to measure *FEV*1%. A patient is forced to breath fully and a spirometer measures volume of the breath. It is a hard work for a patient with lung disease. But computational lung sound analyses don't need a patient's effort and can watch in 24 hours with a small wearable microphone.

1.2 Purpose of Proposed Method

There are method without computational analysis. One of them is *FEV*1% measurement. A little spirometer is used to measure *FEV*1%. A patient blows the machine with effort. It is a hard work for patients with lung disease. But lung sound analysis doesn't need a patient's effort, is a contiguous method and needs only a little wearable microphone on a body of a patient.

1.3 Purpose of Proposed Method

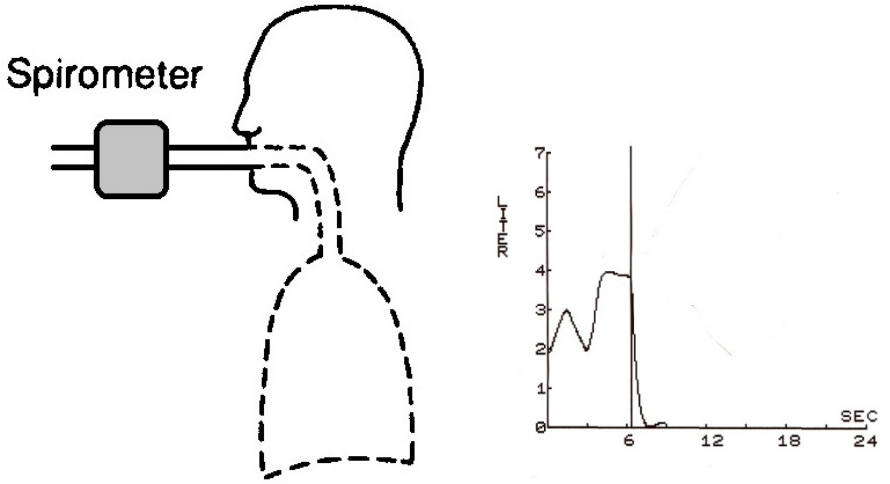
The rest part of this paper goes as below. Section 2 shows sound and nature of a lung. Section 3 shows the detail of the proposed method. Section 4 shows an experimental result. Section 5 shows a consideration. And Section 6 shows a conclusion.

2 Sound and Nature of Lung

2.1 Function of Respiration

Most people concern about their breaths keep their health. There increases number of patients of bronchial asthma, pneumonia or other lung diseases in this decade. A breath shows a condition of an airway and a lung. An airway obstruction often

occurs in asthma or pneumonia. The traditional method to detect airway obstruction is *FEV*1%. It uses a spirometry. It needs a much effort of a patient to blow a spirometry. And a man cannot blow a spirometry in continuity. Once in a day is an ordinary way. But there are many wearable sensors in this days. Lung sounds can be recorded in 24 hours in a day in continuity using one of those sensors. Many works analyzed lung sound.



(a) Usage of Spirometer (b) Graph of Record of Spirometer

Fig. 1 Spirometer

Fig. 1 shows a spirometer. Fig. 1(a) shows usage of a spirometer. A person (a patient or a normal) blows or sucks air from/to his/her lung with effort. The air goes through a tube of a spirometer. There are three measurements *FVC*, *FEV1* and *FEV1%*. *FVC* is Forced Vital Capacity (Liter) the volume of blown air from a full lung condition to an exhausted lung condition with effort. *FEV1* (Liter) is the volume of the air blown in the first 1 second. *FEV1%* is a percentage of *FEV1* per *FVC*.

$$FEV1\% = \frac{FEV1}{FVC} \times 100 \quad (1)$$

2.2 Sounds and Noises of Breath

J. Antônio [2] described lung sounds such as wheeze or else

- Normal Sounds of Lung

1) Tracheal Breath Sounds

The sound occurs at a neck. Mostly its frequency is under 1 kHz.

2) Bronchial Breath Sounds
The sound occurs at thick bronchi. Mostly its frequency is under 1 kHz.

- Adventitious Sound

1) Continuity Adventitious Sound
Rhonchi have thick power between 200 (Hz) and 300 (Hz).
Wheezes have thick poser between over 400 (Hz) and occur in chests.

2) Discontinuous Rale
Coarse crackle has thick power between 200 (Hz) and 500 (Hz).A period of coarse crackle is 15 (ms).
Fine crackle has thick power under 2000(Hz).A period of fine crackle is 5 (ms).

2.3 Medical Approach About Lung Sounds Analysis

Some researchers in medical area analyzed frequency distribution of lung sounds before and after of decrease of *FEV*1 [3-5]. They administered subjects with mild asthma case methacholine or histamine. Administrations of such drugs make bronchial obstructions artificially. R. Beck et al. [3] investigated spectra of sounds with an original *FEV* (i.e. no drug), spectra with *FEV* which are 20% down and spectra with *FEV* which are 40% down. They administered histamine to subject which were from 9 years old to16 years old. They extracted wheeze sounds from the spectra and analyzed them with LSA (Latent Semantic Analysis) algorithm to detect bronchial obstructions. But the result was not so enough. L.P. Malmberg et al. [4] used the highest frequency *Fmax* in a spectrum and the median frequency *F50*. H.J.W. Schreur et al. [5] compared the flow values of breaths measured by a spirometer and LSI (lung sound intensity).

2.4 Information Processing Approach About Lung Sounds Analysis

Some researchers in information processing area worked about lung sounds and lung diseases. R. Palaniappan et al. made an algorithm to classify lung sounds to some classes, wheeze, rhonchi, coarse crackle and fine crackle using machine learning [6-7]. R.J. Riella et al. [8] detected edges of concentrations of wheeze sounds in spectrum and determine whether the sound includes wheeze or not. M. Y. Chen et al. [9] employed back propagation, vector quantization and competitive learning to classify a lung sound to bronchial breath sound, bronchial lung alveoli sound, lung alveoli sounds, wheeze, crackle or stridor. The accuracy of their method is 90% for explicit sound data. M. Bahoura et al. [10] employed wavelet and decrease error detection of wheeze. And also M. Bahoura et al. [11] removed white Gaussian noise using wavelet before to detect wheeze.

3 Method

This section shows algorithms employed by the proposed method in this paper to detect airway obstructions in lungs from lung sounds. We approach another way instead the recent works. The researcher described above wanted to detect individual abnormal sounds in analysis of lung sounds. There are some cases with explicit abnormal lung sounds. But there are other cases without explicit abnormal lung sounds. We focused latter cases. The methods to detect abnormal sounds are not available to the cases without abnormal noises. We found that there were power distributions of sounds between 200Hz and 400Hz which seemed wheeze sounds but they have not explicit peaks in airway obstruction sound records. So we propose a method which is available to no-wheeze cases to detect airway obstructions.

Fig. 2 Electronic Stethoscope

3.1 *Recording of Lung Sound*

We used an electronic stethoscope to record lung sounds as shown in Fig.2.

Table 1 Dimensions of Electronic Stethoscope

Sampling Frequency	4000[Hz]
Recording Mode	Diaphragm
Additive Function	Ambient Noise Canceller

Table 1 shows parameters of an electronic stethoscope when we used it.

3.2 *STFT*

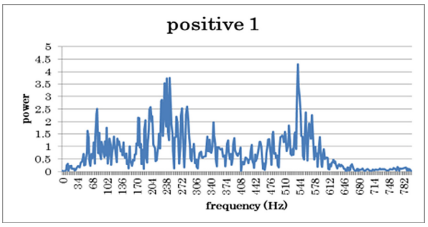
We employ STFT (Short-Time Fourier Transform) as some of recent works described above. We didn't use wavelet and we think that wavelet is one of our future works.

An ordinary Fourier transform is described as an equation shown below.

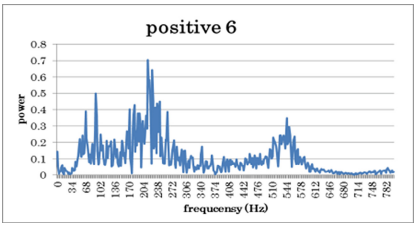
$$F(f) = \int_{-\infty}^{\infty} f(t) e^{-j2\pi ft} dt \tag{2}$$

We select window size 0.124 sec.

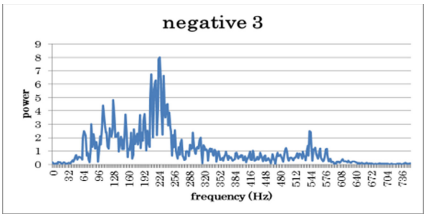
3.3 Power Rate



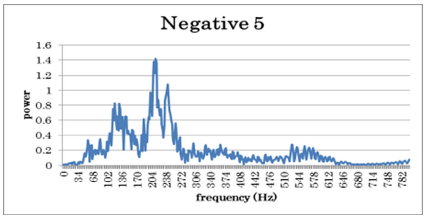
(a) Positive Case of Explicit Wheeze Peak



(b) Positive Case of Small Wheeze Power



(c) Negative Case 1



(d) Negative Case 2

Fig. 3 Frequency Distribution

Fig.3 shows some frequency distribution of power of a lung sound. Fig.3(a) is a typical positive case of airway obstruction. There is a high and precipitous peak near 544Hz. It seemed as a wheeze. And the peak of wheeze is higher than normal sounds which frequency are from 34Hz to 306Hz. Fig.3(b) is a positive case of airway obstruction which peaks over 408Hz are lower than normal sounds. There are many peaks from 408 to 612Hz. We guess that air flow which is a source of wheeze is stable in Fig.3(a). But it is not stable in Fig.3(b) and failed to make a clean sound. Fig.3.(c) shows a negative case. There is a low but precipitous peak

near 544z. It may be detected as a wheeze peak. But power from 416Hz to 782Hz is much smaller than power from 32 to 384Hz. Fig. 3(d) shows also a normal case. There is large power from 34Hz to 374Hz as normal sounds of lung. And power from 408 Hz to 782 Hz was very small. The power of a high area is much smaller than the powers of a low area in negative case. We thought that the power rate of a high frequency rea vs. a low frequency area may reflect the airway obstruction level. The computation of the powers needs not explicit wheeze detection. Equation (3) shows the ratio of power of a high frequency area and power of a low frequency area. $P_{0\text{-}400}$ indicates mean power from 0Hz to 400Hz. $P_{400\text{-}800}$ indicates mean power from 400Hz to 800Hz. We think that the ratio R can be used as an index value to indicate airway obstruction calculated from sounds only without usage of a spirometer.

$$R = \frac{P_{400\text{-}800}}{P_{0\text{-}400}} \tag{3}$$

4 Experimental Result

Fig. 4. shows a distributions of rate R and $FEV1\%$. There are 31 positive cases and 55 negative cases. We measured $FEV1\%$ of a person and record his/her lung sounds. Sometime we administrated the person methacholine to make airway obstruction artificially. The best correlation coefficient was -0.592. We selected 70% as a threshold of $FEV1\%$. A person which $FEV1\%$ is over 70% is negative at airway obstruction. And a person under 70% is positive.

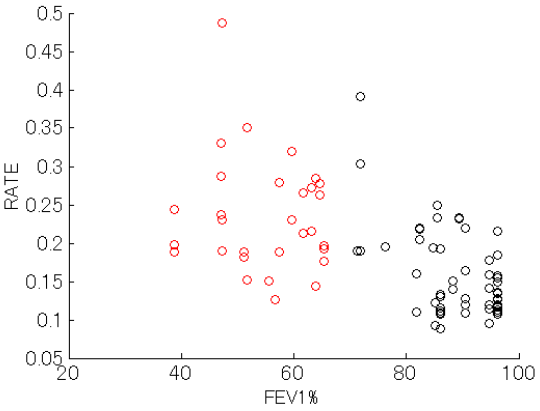


Fig. 4 Rate R vs. $FEV1\%$

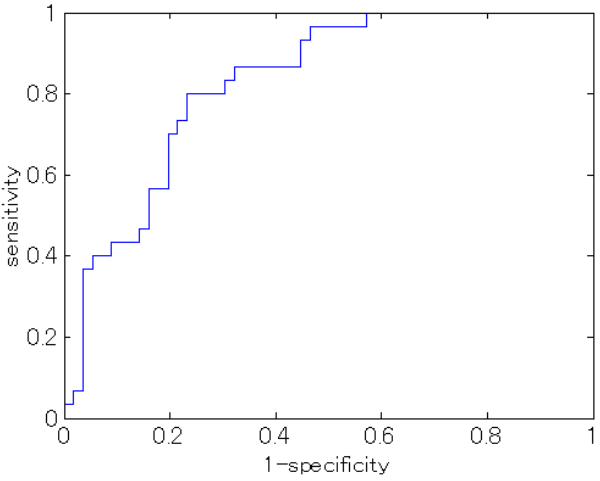


Fig. 5 ROC Curve

Fig.5 shows ROC curve with 70% threshold of *FEV*1% and various thresholds of rate *R*. AUC value is 0.833 . There are 4 values about quality of diagnosis. *TP* is number of "true positive". The person is positive and the diagnosis is positive. *TN* is number of "true negative". The person is negative and the diagnosis is negative. *FP* is number of "false positive". The person is negative and the diagnosis is positive. *FN* is number of "false negative". The person is positive and the diagnosis is negative. A point with a threshold (R=0.181) in Fig.5 indicates sensitivity=0.8 and specificity = 0.78.

$$\text{sensitivity} = \frac{TP}{TP+FN} \tag{4}$$

$$\text{specificity} = \frac{TN}{TN+FP} \tag{5}$$

5 Consideration

The absolute value of correlation coefficient as the result 0.593 is evaluated moderate accuracy. And also AUC value 0.833 is moderate accuracy. A small microphone on a patient skin can easily record lung sounds. An equipment such as a smart phone can be calculate a rate *R* in real time. Then a patient can be monitored about airway obstruction 24 hours in real time. The proposed method may help many of asthma patients.

6 Conclusion

We proposed a new method to detect airway obstruction available also for cases with small power in high frequency area.

In future we will investigate the frequency threshold 400Hz and the use of other function instead of a simple mean.

References

1. Ortiz, G.: Asthma Diagnosis and Management: A Review of the Updated National Asthma Education and Prevention Program Treatment Guidelines. The Internet Journal of Academic Physician Assistants **6**(2) (2008)
2. Antônio, J.: Not all that wheezes is asthma! J Bras. Pneumol. **39**(4), 518–520 (2013)
3. Beck, R., Dickson, U., et al.: Histamine Challenge in Young Children Using Computerized Lung Sounds Analysis. Chest **102**, 759–763 (1992)
4. Malmberg, L.P., Sovijarvi, A.R.A., et al.: Challenges in Frequency Spectra of Breath Sounds During Histamine Challenge Test in Adult Asthmatics and Healthy Control Subjects. Chest **105**, 122–132 (1994)
5. Schreur, H.J.W., Vanderschoot, J., et al.: The effect of methacholine-induced acute airway narrowing on lung sounds in normal and asthmatic subjects. Eur. Respir. J. **8**, 257–265 (1995)
6. Palaniappan, R., Sundaraj, K., Ahamed, N.U.: Machine learning in lung sound analysis. Biocybernetric And Biological Engineering **33**, 129–135 (2013)
7. Palaniappan, R., Sundaraj, K., Ahamed, N.U., Arjunan, A., Sundaraj, S.: Computer-based Respiratry Sound Analysis: A Systematic Review. IETE Technical Review **30**(3), 248–258 (2013)
8. Riella, R.J., Nohama, P., Maia, J.M.: Method for automatic detection of wheezing in lung sounds. Brazilian Journal of Medical and Biological Research **42**, 674–684 (2009)
9. Chen, M.-Y., Chou, C.-H.: Applying Cybernetic Technology to Diagnose Human Pulmonary Sounds. Journal of Medical Systems **38**(6), 1–10 (2014)
10. Bahoura, M., Hubin, M.: Automatic wheeze detection using wavelet packets. In: Conference on Medical and Biological Engineering and Computing VIII Mediterranean, Limassol, Cyprus, pp. 14–17, June 1998
11. Bahoura, M., Lu, X.: Separation of crackles from respiratory sounds using wavelet packet transform. In: The 31st International Conference on Acoustics, Speech, and Signal Processing (ICASSP-06), Toulouse, France, vol. II, pp. 1076–1079, May 2006

Entropy Based Test Cases Reduction Algorithm for User Session Based Testing

Hsu Mon Maung and Kay Thi Win

Abstract Web applications are crucial role for daily user activities such as online banking, online shopping and searching. It is important to ensure the reliability and web application testing has been used in finding various faults in order to improve the quality of reliable web services. Among test cases generation approaches, user session based testing is an approach to create test cases with real user data. However, real user data usage is extremely large and executing all the test cases can be time consuming in practice. This paper describes the test cases reduction approach for analyzing and replaying the large number of test cases generated from user session data. The entropy gain theory is applied in test cases reduction process to get the best test suite that covers all user accesses of web application. To evaluate the effectiveness of proposed method, the analytical results are described in terms of URLs coverage, reduction time and test cases reduction rate.

Keywords User session based testing · Entropy gain theory · Test cases reduction

1 Introduction

As most daily activities rely on the services provided by web applications (WA), the qualities of these applications are central role. Testing web applications is an integral part of software development process in order to ensure software quality. However, web application testing is a very expensive process in terms of time and resources due to the nature of web application. Testing, designing and generating test cases are challenging tasks because web application is complex and changeable. There are different types of web application with the goal of finding faults in the software under development. User session based testing has been recently researched as a way for effectively testing web application. This technique is a capture/replay mechanism that collects user data with user

H.M. Maung(✉) · K.T. Win
University of Computer Studies (UCSM), Mandalay, Myanmar
e-mail: {hsumon77,kthiwin11}@gmail.com

T.T. Zin et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 388,
DOI: 10.1007/978-3-319-23207-2_37

interaction from web server and these collected data are transformed into test cases in the form of http requests. For web application system, field data has the additional advantage because the usage data is independent of the underlying implementation and server technologies, thus reducing the costs of finding inputs [1]. User session based testing is less dependent on fast changing technologies used by web applications and it can generate test cases using real user data without analyzing the internal structure. In this testing, a tester captures user accesses during deployment to create user session which are then replayed as test cases. A major problem with user session based testing is the cost of collecting, analyzing, and replaying the large number of test cases generated from user session data [2]. Many researchers have proposed selection and reduction methods of test suites in user session based testing because the collected daily user logs are million gigabytes that are impossible to replay as test cases. However, the effectiveness of this testing technique depends on the collected user session data set. To design test cases effectively, the strategy is needed to be not only reduce test suite size but also cover each possible user behavior. This paper presents an approach for achieving high efficiency test results in user session based testing by reducing the overhead of selection and analyzing user session data. The main goal of paper selecting test cases for test suite reduction based on entropy value analysis. A reduced set of user session data is produced by applying proposed entropy based reduction algorithm. Some preliminary case studies were carried out to validate the proposed technique and to evaluate its effectiveness.

In the remainder of this paper, the related work is described in Section 2. Testing web applications and user session based testing are described in Section 3 and 4 respectively. In Section 5, we present methodology, entropy based heuristic for test suite reduction. Section 6 describes experimental study with two subject applications for evaluation process. The paper is concluded in Section 7.

2 Related Work

In this section, several researches related with testing web application are described.

Sampath et al. [3] explored the possibility of using concept analysis for achieving reduction and scalability in user session based testing of web applications. This method is completely automated user session selecting, reduction through replay process. The studies showed that concept analysis can provide incrementally updating reduced test suite. The authors also admitted the importance of request data and ordering.

The studies [4] explored a method of estimating dependencies automatically and using them to arrange the test suite. The authors depicted some limitations of an approach to testing Web applications automatically and introduce some ideas for improving upon it.

Ebrahim Shamsoddin-Motlagh reported a survey of recent research to generate test case automatically. Those are presented from UML based, graph based, formal methods, web application, web service, and combined methods [5].

H.M.Maung [6] proposed the framework for user session data reduction in web application testing. The authors also discussed validation methods for evaluating the effectiveness of this approach.

The studies [7] explored a method of estimating dependencies automatically and using them to arrange the test suite. The authors depicted some limitations of an approach to testing Web applications automatically and introduce some ideas for improving upon it.

3 User Session Based Testing

In user session based testing, each user session is a sequence of user requests in the form of baser requests and name-value pairs [11]. It generates test cases that can effectively detect residential faults with the use of field data. The key advantage is that the minimal configuration changes need to be made to the web server to collect user requests [12]. In addition, user session based testing is useful when the program specification s and requirements are not available for test cases generation [3]. A user session is transformed into a test case by changing each logged request into HTTP format that can be sent to the server. When a request from a new IP address arrives at the server, a user session is identified as initial and when the user leaves or session time out, the user session is identified as the end. A time out interval of 30 minutes is taken to identify the user session.

4 Test Cases Reduction Technique

The number of user session data can be very large due to frequently usage of web applications. Using all of the user sessions as test cases may not be practical when testing the application. One of the web application natures, frequently maintenance changes can cause some test cases to become obsolete. In addition, redundant user requests are also contained in user session data and it may be overload for generating as test cases. Therefore, test suites reduction technique is needed with the criteria of covering all base requests in original test suite. Test suite reduction is a test suite management method where a smaller set of test cases are selected from a large original suite while maintaining the requirement coverage of the original suite [8]. Test requirements coverage is very important and well accepted measure for deciding when to stop testing, selecting test cases and reducing test suites [9].

5 Methodology

In this paper, new heuristic for test suite reduction is proposed by applying entropy gain theory. The core definition of information theory is the entropy, a measure for uncertainty of a random variable [13]. Shanon's information entropy can be expressed as following equation and assume that N objects and a variable containing K categories [10]:

$$H = -\sum_{i=0}^{n} P_i \log P_i \tag{1}$$

where H is entropy and pi is the probability of being in category *i* out of the K possible categories. The values of H vary from the minimum value of zero to a maximum value of log K (K is the number of categories). When a single category is occupied by all N objects, the value of H is minimum value of zero and when all objects are evenly distributed among the K categories, the maximum value of H is attained.

In this paper, the entropy based reduction method is proposed to reduce test cases for user session based testing. By analyzing the entropy values, the proposed concept is that the higher entropy value leads to the more URLs covered. However, to normalize the entropy value to the range [0, 1], the base logarithm is needed to choose the total number of links. Therefore, the entropy equation becomes as follows:

$$H = -\sum_{i=0}^{n} P_i \log_K P_i \tag{2}$$

where K is the total number of links of web application that are extracted by using link extraction tool. The goal of this heuristic is to reduce the original user sessions into an equivalent smaller one with the full requirements coverage. Therefore, the reduction algorithm is proposed by applying entropy gain theory on user session data. Our reduction algorithm, shown in Table 1, use Shannon's entropy gain function for analyzing coverage of user accessed based requests.

Table 1 Entropy based Test Cases Reduction Algorithm

Algorithm. Test Cases Reduction

Input: user sessions U = (u_1,u_2,…,u_m) with user session u_i consisting of n requests r_1,r_2,…,r_n
Output: Reduced Test Suite T = (t_0, t_1, …, t_n)
- Calculate entropy value E1,E2,…,Em for each user U1,U2,…,Um in user session.
- T = max ({E(u1), E(u2),….,E(um)}) as test case ti
- Request R = ri € ti
- Select next highest entropy E (ui) as test cases tj (i≠j)
if (rj € R) then
Remove ui and select next highest entropy E (ui)
else
ui is selected and add it to the reduced test suite T
-Repeat the process until all base requests are satisfied in original test suite

To demonstrate proposed reduction algorithm, there are three main components as illustrated in Fig. 1. The first step, user session data collection is easily accomplished by capturing data from web server. These access logs are converted into test cases in second step and finally, some heuristics are applied in reducing phase.

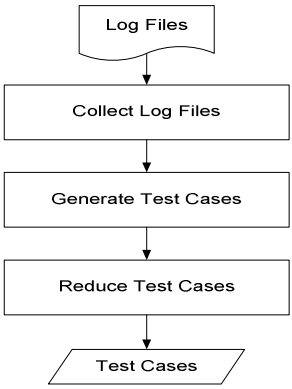


Fig. 1 Framework of Proposed System

5.1 User Session Collection and Test Cases Generation

In the first step of proposed system, the user sessions data of specific web application system as in Fig. 2 are collected. These access logs are collected from Digital Library system (DLS) and the official website http://www.ucsm.edu.mm of our university. The user access logs are parsed into different fields and we remove the unnecessary data from these logs such as .jpg.

As a second step, these access logs are converted into test cases in the form of http requests that can be sent to the server. A user session based test case is a sequence of HTTP request (GET or POST) containing base requests and name value pairs that are recorded when a user accesses the application [6]. Many researchers have proposed different strategies for construction test cases from user sessions [3],[4],[5],[6].

```
192.168.1.87 - - [23/Jan/2014:14:25:51 +0630] "GET /E-
Library/index.php/books/text-books/second-year HTTP/1.1" 200
89438 "http://192.168.1.7/E-Library/" "Mozilla/5.0 (Windows NT
6.0; rv:26.0) Gecko/20100101 Firefox/26.0"
192.168.1.87 - - [23/Jan/2014:14:26:01 +0630] "GET /E-
Library/index.php/books/lecturer-slide/second-year HTTP/1.1" 200
14606 "http://192.168.1.7/E-Library/" "Mozilla/5.0 (Windows NT
6.0; rv:26.0) Gecko/20100101 Firefox/26.0"
192.168.1.87 - - [23/Jan/2014:14:26:03 +0630] "GET /E-
Library/media/system/images/printButton.png HTTP/1.1" 200 228
"http://192.168.1.7/E-Library/index.php/books/lecturer-
slide/second-year" "Mozilla/5.0 (Windows NT 6.0; rv:26.0)
Gecko/20100101 Firefox/26.0"
192.168.1.87 - - [23/Jan/2014:14:26:03 +0630] "GET /E-
Library/media/system/images/emailButton.png HTTP/1.1" 200 277
"http://192.168.1.7/E-Library/index.php/books/lecturer-
slide/second-year" "Mozilla/5.0 (Windows NT 6.0; rv:26.0)
Gecko/20100101 Firefox/26.0"
192.168.1.87 - - [23/Jan/2014:14:26:45 +0630] "GET /E-
Library/index.php/books/lecturer-slide/second-year HTTP/1.1" 200
14606 "http://192.168.1.7/E-Library/index.php/books/lecturer-
slide/second-year" "Mozilla/5.0 (Windows NT 6.0; rv:26.0)
Gecko/20100101 Firefox/26.0"
192.168.1.87 - - [23/Jan/2014:14:26:57 +0630] "GET /E-
Library/index.php/timetables/master HTTP/1.1" 200 14549
"http://192.168.1.7/E-Library/index.php/books/lecturer-
slide/second-year" "Mozilla/5.0 (Windows NT 6.0; rv:26.0)
Gecko/20100101 Firefox/26.0"
192.168.1.87 - - [23/Jan/2014:14:27:10 +0630] "GET /E-
Library/index.php/timetables/second-year HTTP/1.1" 200 78826
"http://192.168.1.7/E-Library/index.php/timetables/master"
```

Fig. 2 Sample Access Log

5.2 User Session Data Reduction

Before reduction process, we need to identify which test requirements are used for coverage criteria. The different test requirements from user session based test cases are generally defined as base, sequence, sequence name, name, and name_vlue [9]. In our proposed system, base request coverage (base) is used as coverage criteria and thus base form of requests are extracted in user session data.

In reduction process, we need to calculate the entropy value of each user in user sessions pool. We select the highest entropy value of user is selected as a test case because it covers all or most URLs in web application than others. According to table 2, entropy value of user 3 is highest and it is selected as a test case. All requests in selected test case are marked. For example, numbers of base requests covered by U3 are 5 on total numbers of base requests 6. We select the next highest entropy value of user and compare the requests with selected test case to check whether all base requests are covered or not. In some cases, there are two or more users that have same entropy values mean that the amount of base requests accessed by user are the same. In this case, we need to consider the requests contained in these users.

According to our example in Table 2, the entropy values of user 5 and user 6 are the same. By comparing the requests in these users, we found that the base request (view.php) does not include in selected test case U3 and all base requests in U6 are also contained in test case U3. Therefore, selecting only two test cases U3 and U5 can cover all base requests in original test cases instead of using six test cases. We need to continue these processes until all of the base requests in original test suites are satisfied. Our entropy based test cases reduction algorithm shown in Table 1 enables not only reduce test suite size but also cover base requests as original one. By reducing test suite size, the testers save time because the large original test suite does not need to execute.

Table 2 Sample Entropy Value Table of Each User

User	Home. php	Search. php	Books. php	Login. php	View. php	Ebooks. php	Entropy Value
U1	1	1	1	0	0	0	0.61314
U2	1	1	0	1	0	0	0.61314
U3	1	1	1	1	0	1	**0.89824**
U4	1	1	1	0	0	0	0.61314
U5	1	1	1	0	1	0	0.77371
U6	1	1	0	1	0	1	0.77371

6 Experimental Framework

6.1 Subject Applications

To validate the proposed reduction algorithm, some case studies were carried out in this section. The results of case study involving two real web applications (Digital Library System and www.ucsm.edu.mm) will be presented. Digital Library System enalbes all of the students to access e-books, syllablus, old questions and update timetables of each class. The 222 user sessions are collected from web server by the student users. The requests to administraor are removed because administrative fuctionalities are not considered in our study. Our official site, ucsm.edu.mm, is developed by our e-government team for students where we collected about 1000 user sessons. Users can view proceedings and workshop announcement, exam results, lecture invitation, and activities and other related information in this site.

6.2 Reduction Rate and Coverage Analysis

In our experiment, the effectiveness of our proposed reduction strategy was studied by evaluating the base requests coverage, reduction rate and reduction time. In terms of base request coverage as described in equation 3, we obtained full coverage rate (100%) because the reduced test cases can cover all base requests in original test suite. The proposed reduction algorithm stops reduction process when all requests in original test suite are covered. According to experimental results, our test cases reduction algorithm produces a reduced test suite that is smaller in size without loosing requests in the user session data.

$$base\ request\ coverage = \frac{no.of\ base\ \ requests\ \ in\ \ reduced\ test\ suite}{no.of\ base\ \ requests\ \ in\ original\ test\ suite} * 100\ \% \qquad (3)$$

To evaluate the effectiveness in reduced size of test suite, the reduction rate of proposed algorithm is calculated using user session data from both applications DLS (Digital Library System) and the web site (ucsm.edu.mm). We found that if there are many users that have high entropy values in user sessions, the algorithm can reduce test cases significantly. Table 3 presents the reduction rate of proposed method. These test cases reduction rates depend on the usage of base requests by user in user sessions pool. The results show that our entropy based reduction approach reduced over 90% test cases on both applications. This outcome is fine for user session based testing with the goal of reducing cost in testing process.

Table 3 Reduction Rate of Proposed Method

Applications	Original Test Suite	Reduce Test Suite	Reduction Rate
DLS	222	20	90.99 %
ucsm.edu.mm	277	18	93.50 %

6.3 Reduction Time Computation

During testing to web application, a primary concern is time taken to execute the test suite. Test cases reduction time is important fact because it affects the effectiveness of reduction techniques. Factors that affect the test cases reduction time are complexity of the requirements used by the criteria and the algorithm implementation [8]. Therefore, the time taken in test cases reduction process by applying proposed algorithm is measured to evaluate test effectivnenss.

Table 4 Reduction Time of Test Cases

Log Files	Original Test Suite	Reduce Test Suite	Reduction Time (ms)
Access Log 1	161	15	28
Access Log 2	277	18	26
Access Log 3	276	22	60

In this work, we use three user access logs of ucsm.edu.mm to analyze the reduction time. Table 4 shows the reduction time (ms) for the test cases of our subject applications. We divide the user access logs into three access logs based on access time that is january 2013 to march 2013 be access log1, april 2013 to june 2013 be access log2 etc. Different number of users are included in these access log files. We observed that the variation of reduction time by each access log depend on number of users contained in user session and entropy values of users which means that usage of base requests by users. From Table 4, we noted that the reduction time for access log 3 is longer than access log 2 although the more users are contained in access log 2. In this case, we found more users who have same entropy value in access log 3 than access log 2. Because we select test cases based on entropy values, we need to check repeatly whether base requests are equal or not in same entropy cases. Therefore, we supposed that the more same entropy cases, the longer reduction time taken by proposed algorithm.

7 Conclusion

In user session based testing of web system, web usage logs are very large and thus resulting in a large set of tests cases. Using large amount of test cases in testing is not practical within time constraint. In this paper, the new reduction approach is proposed to reduce test suite with the full base requests coverage. We have presented test cases reduction algorithm based on entropy value analysis to satisfy all base requests as original test suite. The empirical results of our approach show that the reduced test cases can cover all base requests in terms of coverage criterion. Because proposed method uses entropy values to decide requests coverage, the test cases can be reduced without over-reducing user session data. In this paper, reduction rate, time and base request coverage are presented for

evaluating the results of proposed approach. We have not yet fully compared our approach to current user session based testing techniques. In the future, the abilities of fault detection will be evaluated with other test cases reduction approaches.

References

1. Sprenkle, S., Gibson, E., Sampth, S., Pollock, L.: A case study of automatically creating test suites from web application field data. In: TAVWEB 2006, Portland, Maine, USA, July 2006
2. Sprenkle, S., Gibson, E., Sampth, S., Pollock, L.: An empirical comparison of test suite reduction techniques for user session based testing of web application. In: ICSM 2005: Proceeding of the 21st IEEE Int. Conf. Softw. Mainten., Budapest, Hungary, pp. 587–596 (2005)
3. Sampth, S., Mihaylov, V., Souter, A., Pollock, L.: A scalable approach to user session based testing of web applications through concept analysis. In: ASE 2004: Proc.19th Int. Conf. Automated Sofw. Eng., Washington, DC, USA, pp. 132–141 (2006)
4. Pobletts, A., Cobb, C., Simko, L.: Toward an Effective Data Model and User Session Dependency Model. In: International Conf. (2011)
5. Sampath, S., Bryce, R.C.: Improving the Effectiveness of Test Suite Reduction for User Session based Testing of Web Application. Information and Software Technology **54**, 724–738 (2012)
6. Maung, H.M.: Test Case Reduction Approach in User Session based Testing for Web application. In: ICCA, February 2013
7. Sampth, S., Bryce, R.C.: Prioritizating User Session based Test Cases for Web application
8. Sampath, S., Bryce, R.C.: Improving the Effectiveness of Test Suite Reduction for User Session based Testing of Web Application. Information and Software Technology **54**, 724–738 (2012)
9. Sprenkle, S.E.: Strategies for Automatically Exposing Faults in Web Applications. University of Delaware (2007)
10. Bailey, K.D.: Entropy System Theory, University of California, Los Angeles, USA
11. Sampath, S., et al.: Applying Concept Analysis to User Session based Testing of Web Applications. IEEE Transactions on Software Engineering **33**(10), October 2007
12. Maung, H.M.: An efficient test cases reduction approach in user session based testing. In: 6th International Conference on Education Technology and Computer (ICETC) 2014, Singapore (2014)
13. Herbold, S., Karlshafen, B.: Usage-based Testing for Event Driven Software (2012)

Part V
Text Analysis Technologies and Development Strategies for e-Learning

Fusion of E-Textbooks, Learning Management Systems, and Social Networking Sites: A Mash-Up Development

Masumi Hori, Seishi Ono, Shinzo Kobayashi, Kazutsuna Yamaji, Toshihiro Kita and Tsuneo Yamada

Abstract Online education has provided good opportunities for educationally disadvantaged people. However, some traditional learning management systems (LMSs), the base systems of online education, had the limitations in offering standardized education for diversified learners with different skills, objectives, abilities, preferences, and backgrounds. In addition, the traditional LMSs, which required a constant connection of the Internet, could not be used where it is not available, that is, in the half of the world. Thus, we developed a new learning platform for large-scale online courses (LSOC), called "the Creative Higher Education with Learning Object (CHiLO)". CHiLO is a comprehensive, open-network learning system which can realize e-textbooks, competency-based education (CBE), digital badges, and social learning. CHiLO can contribute to future

M. Hori(✉) · S. Ono
NPO CCC-TIES, Nara, Japan
e-mail: {hori,ono}@cccties.org

S. Kobayashi
SmileNC & Co, Saitama, Japan
e-mail: shinzo@smilenc.com

K. Yamaji
National Institute of Informatics, Tokyo, Japan
e-mail: yamaji@nii.ac.jp

T. Kita
Kumamoto University, Kumamoto, Japan
e-mail: t-kita@kumamoto-u.ac.jp

T. Yamada
The Open University of Japan, Chiba, Japan
e-mail: tsyamada@ouj.ac.jp

T.T. Zin et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 388,
DOI: 10.1007/978-3-319-23207-2_38

research on next-generation learning content based on e-books and a flexible, diversified learning environment for people worldwide.

Keywords Large-Scale Online Courses (LSOC) · E-textbook · Competency-Based Education (CBE) · Digital badges · Social learning

1 Introduction

Education for All (EFA) is a large project and major challenge issued by the United Nations Educational, Scientific and Cultural Organization (UNESCO). This movement involves a global commitment to provide basic quality education for all children, youth, and adults (see http://www.unesco.org/new/en/education/themes/leading-the-international-agenda/education-for-all/). However, accomplishing EFA's goal through traditional teaching methods—e.g., building a massive number of brick-and-mortar schoolrooms and supplying many teachers to educate the world's masses—is not realistic. Online learning can disseminate education worldwide at a low cost. In fact, massive open online courses (MOOCs), a type of large-scale online course, can radically contribute to enhancing opportunities for higher education around the world.

However, MOOCs incorporate potential issues found in traditional learning management systems (LMS). Traditional LMSs disregard two types of diversity observed in large-scale online courses (LSOCs); the standardization of education in a learning environment including diverse students and the geographical divide in the digital environment.

Considering these issues, we have developed the Creative Higher Education with Learning Object (CHiLO) learning platform. CHiLO aims to provide a device-agnostic, ubiquitous learning environment through e-textbooks. It possesses effectively high portability in the electronic publication 3.0 (EPUB3) format and a comprehensive, open-network learning system created by combining various existing technologies, such as LMSs and learning resources, which includes open educational resources (OER) in open-network communities such as social networking services (SNSs).

In this study, we report on CHiLO's architecture and its possibilities based on some experimental results.

2 The LMS in Large-Scale Online Courses

2.1 Standardized Education in Large-Scale Online Courses

In LSOCs, learners have different skills, objectives, abilities, preferences, and backgrounds. Despite that, a traditional LMS offers all learners the same content during the same term and the same assessments with the same criteria (Mintz 2014, Mazoue 2013, Wilkowski et al. 2014). The post-MOOC movement observed after MOOCs gained popularity in 2012 seemed to struggle with standardized education.

- Competency-based education (CBE) focuses on effective learning for adults, e.g., working and self-supporting students, over a short amount of time (Sturgis et al. 2011).
- The Task Force on the Future of Massachusetts Institute of Technology (MIT) Education provided further insights into unbundling education, which involves using different roles—such as classrooms, labs, and mentoring—as modules. A module is defined by its corresponding outcomes, e.g., its instruction and assessment. Each module is re-bundled with competency-based assessments or new assessment methods relating directly to measurable outcomes for a class or module (Force 2013).
- The Nanodegree (see https://www.udacity.com/nanodegree), conferred by Udacity, a for-profit educational organization with a MOOC platform, provides learners with bite-sized bundles of knowledge and immediate motivation for acquiring a degree. Furthermore, its curriculum is designed for acquiring specific business skills over 6 to 12 months (10–20 hours/week) for $200 a month (Porter 2014).

2.2 Geographical Digital Divide

Another challenge of traditional LMSs is the geographical digital divide. Most LMSs are based on web services requiring Internet access. However, about 60% of all people globally do not have Internet access (ITU 2015). Furthermore, 80% globally do not have personal computers (The World Bank 2012). Therefore, online learning that requires a constant Internet connection is unavailable to them.

In contrast, mobile communication devices are ubiquitous. The International Telecommunication Union(ITU) stated, "Globally, mobile-broadband penetration will reach 32% by the end of 2014—almost double the penetration rate just three years earlier (2011) and four times as high as five years earlier (2009)" (2015). Mobile communication devices that provide satellite communication and a personal area network (PAN), such as Bluetooth, which offers a traditional telephone infrastructure with Internet access, have proliferated worldwide even in areas without regular Internet access. Therefore, the use of mobile devices could provide a solution to these challenges.

Nevertheless, mobile devices present different challenges (Deb 2012). Mobile devices' essential problems are their small screens, lack of keyboards, network speed, reliability, short battery life, and limited content and software applications.

3 The Architecture of CHiLO

3.1 CHiLO's Technology Components

CHiLO provides flexible, diversified service for online learning based on various computer network environments, devices, learners' skills with e-books, CBE, digital badges, and social learning (see Fig. 1).

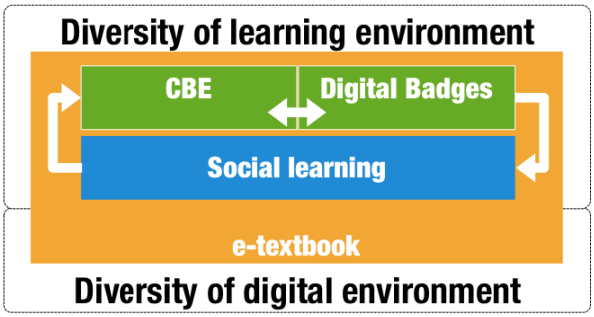


Fig. 1 CHiLO's technology components

CBE, Digital Badges, and Social Learning

We can expect to provide a flexible approach for diverse learning environments by combining CBE, digital badges, and SNS.

The CBE approach provides short-term learning content based on competencies according to the learner's skills, objectives, abilities, preferences, and background. In addition, learners can demonstrate acquired skills and knowledge in CBE by using digital badges. Digital badges, which are part of a conventional system that has been successful in motivating participants by showcasing challenges, have been used in SNSs such as FourSquare and Gamification (Deterding 2012). CBE using digital badges is focused on micro-credentials, which could be connected with learning paths in current efforts to demonstrate learners' mastery of those competencies (Sturgis at al. 2011).

Many studies strongly suggest that cooperative learning is more effective than individualistic learning in contributing to motivation, raising achievement, and producing positive social outcomes (Chen and Bryer 2012). In LSOCs, learners do not learn from a tutor but rather learn on their own with learning materials. A learner who earns digital badges frequently exchanges information with other learners in their communities as a "connoisseur" (see "The CHiLO community" below). In this way, social learning provides the functions of discovering, sharing, aggregating, and repurposing.

e-textbook

With the advent of the EPUB3 format, which offers greater sourcing flexibility, e-books can now include media-rich, interactive content. In one package, an e-textbook can contain all the resources a student needs (Smith and Kukulska-Hulme 2012). Therefore, we can expect to close the geographical digital divide by bundling CBE, digital badges, and social learning into e-books.

Developed by the International Digital Publishing Forum (IDPF), EPUB3 is a distribution and interchange format standard for e-books (Polanka 2013). EPUB3 is not only device-independent but also available on- or offline so that a file can be opened on a PC, tablet PC, or mobile device as long as an e-book reader application is installed.

In education, tutors can easily repurpose and adapt learning materials offered in the EPUB3 format to improve learning outcomes. In addition, they offer a way to avoid vendor lock-in (Belfanti 2014). New generation e-textbooks, such as EDUPUB and EPUB3, show equal or greater educational effects than traditional LMSs. Furthermore, e-textbooks offer unique advantages for distance-education students as well as situational reading (Smith and Kukulska-Hulme 2012). Thus, they have the potential to not only close the geographical digital divide but also provide diverse learning environments.

3.2 CHiLO's Learning Components

CHiLO, which is based on e-textbooks, aims to offer an affordable, scalable design with regard to LSOCs. It consists of the following six components (see Fig. 2):

- CHiLO books adopt a "mash-up" approach to e-textbooks in the EPUB3 format.
- CHiLO lectures are based on one-minute nano lectures embedded in CHiLO books.
- CHiLO badges in CHiLO books provide authentication and certification using Mozilla Open Badges (see http://openbadges.org).
- CHiLO communities offer learning communities built in social networking services, bulletin boards, and chat rooms.
- CHiLO analytics recommends a learning environment suited to the individual learner.
- The CHiLO reader is the dedicated CHiLO book-browsing software program.

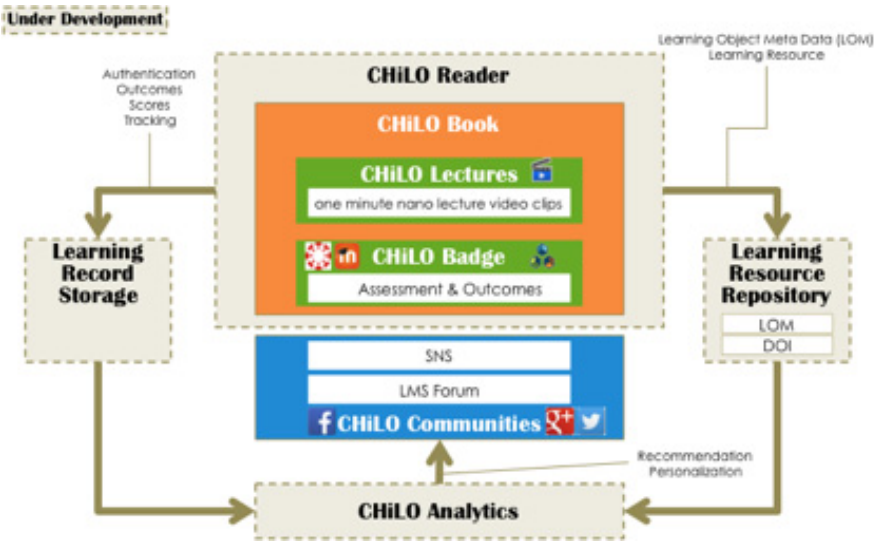


Fig. 2 CHiLO's Component

The CHiLO Book

CHiLO books, a core CHiLO component created using EPUB3 e-textbooks, contain media-rich content, including graphics, animation, audio, and embedded video. Based on the micro-credential method, CHiLO books consist of the learning materials used during a classroom hour. Those who complete a CHiLO book receive a CHiLO badge as a certificate of completion.

The CHiLO Lecture

CHiLO lectures contain videos with scripts, quizzes, and other learning materials. Videos offer one-minute nano lectures. This concept emerged from an experiment revealing that online learners' average viewing time is approximately one minute (Hori et al. 2013). A CHiLO lecture is equivalent to one section in a traditional textbook. A CHiLO book contains approximately 10 CHiLO lectures, and a standard CHiLO course, which is comparable to a traditional university course with one academic credit, consists of 10 CHiLO books.

The CHiLO Badge

Performing indirect assessments, e.g., of learning time and academic workload, is difficult in LSOCs. Although CHiLO adopts a direct-assessment approach for learning outcomes, completion of a CHiLO course is measured in standard course hours corresponding to academic credits.

Whenever a learner completes a CHiLO book, he or she receives a CHiLO badge, which is a simple mechanism to measure successful outcomes in CHiLO. When a tutor wishes to check a learner's progress, the tutor asks the learner to present CHiLO badges, thus removing the need to confirm using indirect assessment tools, such as grade books or tracking past results or test scores.

The CHiLO Community

The CHiLO community provides a social network function. Learners may share a downloaded a CHiLO book and have discussions on an open SNS on the Web, e.g., Facebook and Twitter.

The CHiLO community is comprised of many learners and a few tutors, known as "connoisseurs." These tutors act as substitutes for teachers. A learner who has studied and completed CHiLO books in a specific field can became a connoisseur. The connoisseur and learner are on equal footing, so the connoisseur often exchanges information with learners in the community.

In the CHiLO community, a learner does not learn from a tutor but rather learns independently using CHiLO books as a learning resource. In this way, learners are constantly required to find suitable CHiLO books within the community. The CHiLO community provides methods for discovering, sharing, aggregating, and repurposing CHiLO books for learners.

The CHiLO Analytics
CHiLO analytics recommend learning content, learning methods, and a learning community that fits the individual learner's purposes and preferences. This is made possible by analyzing the learner's activity logs, which are stored in the learning record storage (LRS), and the Institute of Electrical and Electronics Engineers (IEEE) Learning Object Metadata (LOM) in the learning resource repository.

The CHiLO Reader
The CHiLO reader is an e-book reader application optimized for CHiLO books. Its purpose is to enhance the usability of CHiLO books. The CHiLO reader is compatible with three types of CHiLO books: embedded, EPUB3-based, and web-based. The CHiLO reader also records learning history (outcomes, scores, tracking, etc.) when it is offline and sends the history to a learning record storage (LRS) when it goes online.

4 Demonstration Experiments

4.1 Experimental Methodology

In collaboration with the Open University of Japan (OUJ) and the Japan Foundation, we produced 10 CHiLO books titled "Nihongo Starter A1 (NSA1)," which include 10 successive lessons for those learning Japanese for the first time.

As a demonstration experiment, we distributed two types of CHiLO books—EPUB3-based and Web-based—from the NSA1 series over approximately one year (April 2014 to March 2015) at no charge through three different distribution channels (Table 1). One of the distribution channels, the OUJ-MOOC site, is a platform supported by JMOOC, a MOOC provider in Japan (see http://www.jmooc.jp/en/about/).

Table 1 Distribution channel

Distribution channel	EPUB3-based	Web-based
OUJ-MOOC site	✓	✓
iBooks Store	✓	N/A
Google Play	✓	N/A

4.2 Results

The results of the demonstration experiment are listed below:

- In all, 17,590 EPUB3-based CHiLO books and 5260 web-based CHiLO books were downloaded in 104 countries using the three distribution channels.
- Comparing the number of Lesson 1 downloads (6774 books) to Lesson 10 downloads (1304 books), Lesson 10 downloads were only 19% of those for Lesson 1.

- Of the learners, 3156 took assessment examinations at least once. Of the learners who took the online test for Lesson 1 (786 people), 18% (145 people) went on to complete all 10 lessons and earned 10 badges.
- In all, 3181 learners participated in the CHiLO communities (Classes 1–5) on Facebook. Further, learners made 1219 posts in the communities, posted 4046 comments, and gave 5808 "likes."." Participants posted about their positive experiences and showed off the badges they had earned. Furthermore, participants who had completed the course tended to provide helpful suggestions to participants who were still taking the course.
- With regard to the analysis of device-specific access to the Moodle quiz module, we divided access logs into EPUB3-based and web-based CHiLO books; for web-based books, about 69% of all learners' access was from PCs; for EPUB3-based books, approximately 73% of all learners' access was from mobile devices, such as smartphones and tablet PCs. The questionnaire results from those who earned badges in this demonstration experiment (n = 105) showed that 50.5% (53) of the respondents used EPUB3-based CHiLO books in some way (see Fig. 3).

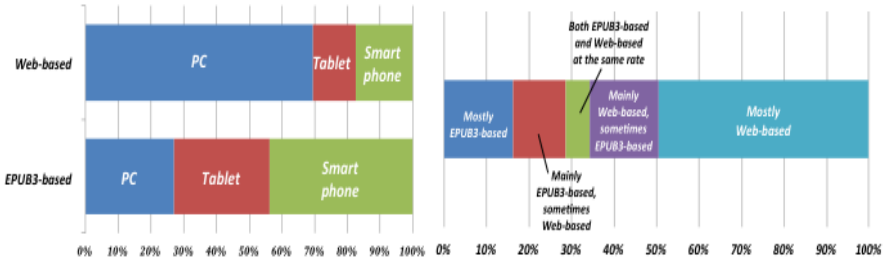


Fig. 3 Ratio of people who responded to the questionnaire regarding CHiLO book format and devices (left). Questionnaire results: Which CHiLO book did you use, EPUB3-based or web-based? (right)

5 Discussion

5.1 Completion Rate

We obtained positive results showing that 18% of the learners who attempted the Lesson 1 examination completed the course. This result is good, considering that the typical completion rate in MOOCs is said to be less than 10%. However, a rigid comparison is not possible because learners did not have to declare enrollment when they began learning in this pilot study.

5.2 Traditional e-book Reader

Based on the demonstration experiment results, we found that a kind of mutual learning occurred in the learning community: Learners who had completed the course tended to provide helpful suggestions to learners who were still taking the course. CHiLO seemed to cultivate each learner's individuality, as opposed to standardized education.

With regard to geographical digital divide issues, CHiLO offers affordable formats for people in 104 countries, including those in rural countries. In addition, learners selected web-based or EPUB3-based CHiLO books and chose to use the books on PCs or smartphones according to their preferences and lifestyles. This result demonstrates that CHiLO could provide flexible, diverse learning environments that are also device-independent, network-independent, and anytime–anywhere.

In the experiment, learners reported that video lectures did not play or that they could not access some quizzes at the ends of chapters in the EPUB3-based CHiLO books. Most e-book readers do not support embedding videos that meet requirement specifications of EPUB3. Quizzes written in JavaScript or JSON do not work in many e-book readers in an offline environment. This experiment could not be conducted with CHiLO analytics or the CHiLO reader, but these technical difficulties associated with traditional e-book readers are expected to be resolved soon.

6 Conclusion

A kind of mutual learning occurred in the learning community, thus addressing some challenges of standardized education. Learners who had completed the course tended to provide helpful suggestions to learners who were still taking the course. In addition, Spanish-speaking learners volunteered to form a learning group in which they translated the NS A1 learning materials into Spanish. The CHiLO seems to cultivate each learner's individuality, as opposed to standardized education.

E-textbooks, such as CHiLO books, are now being introduced into education, and their improvement has been widely studied. The IDPF has proposed that the EDUPUB format meet next-generation learning-content requirements based on the EPUB3 format (see http://www.idpf.org/epub/profiles/edu/spec/). However, the implementation of these books is still being discussed. Our study is meaningful to not only future research on next-generation learning content based on e-books but also efforts to offer flexible, diversified learning environments for people all over the world.

References

Belfanti, P.: What is EDUPUB." EDUPUB 2 Workshop Salt Lake City (2014). http://www.imsglobal.org/edupub/WhatisEdupubBelfantiGylling.pdf (accessed April 20, 2015)

Chen, B., Bryer, T.: Investigating instructional strategies for using social media in formal and informal learning. The International Review of Research in Open and Distributed Learning **13**, 87–104 (2012)

Deb, S.: Multimedia Technology and Distance Learning Using Mobile Technology in Developing Countries. INTECH Open Access Publisher (2012)

Deterding, S.: Gamification: designing for motivation. Interactions **19**(4), 14–17 (2012)

Force, M.T.: Institute-wide Task Force on the Future of MIT Education: Preliminary Report. Future of MIT Education (2013). (accessed February 17, 2014)

ITU The World in 2014: ICT facts and figures. Tech. Rep. (2014). https://www.itu.int/en/ITUD/Statistics/Documents/facts/ICTFactsFigures2014-e.pdf (accessed April 20, 2015)

Hori, M., Ono, S., Yamaji, K., Kobayashi, S., Kita, T.: One-on-one approach for open online courses focusing on large-scale online courses. In: Proceedings of the 5th International Conference on Computer Supported Education, pp. 177–182 (2013)

Mazoue, J.G.: The MOOC model: challenging traditional education (2013). http://www.educause.edu/ero/article/mooc-model-challenging-traditional-education (accessed April 20, 2015)

Mintz, S.: The Future of Higher Education. Inside Higher Ed. (2014). https://www.insidehighered.com/blogs/higher-ed-beta/future-higher-education (accessed April, 20 2015)

Polanka, S.: What Librarians Need to Know About EPUB3 (2013). http://corescholar.libraries.wright.edu/ul_pub/159/ (accessed April 20, 2015)

Porter, E.: A Smart Way to Skip College in Pursuit of a Job Udacity-AT&T 'NanoDegree' Offers an Entry-Level Approach to College. The New York Times Economy (2014). http://www.nytimes.com/2014/06/18/business/economy/udacity-att-nanodegree-offers-an-entry-level-approach-to-college.html (accessed April 20, 2015)

Sturgis, C., Patrick, S., Pittenger, L.: It's not a matter of time: highlights from the 2011 competency-based summit. International Association for K-12 Online Learning (2011)

Smith, M., Kukulska-Hulme, A.: Building mobile learning capacity in higher education: E-books and iPads. In: Marcus, S., Jari, M., Mike, S. (eds.) 11th World Conference on Mobile and Contextual Learning, pp. 298–301. Helsinki: CELSTEC & CICERO Learning (2012)

Wilkowski, J., Deutsch, A., Russell, D.M.: Student skill and goal achievement in the mapping with google MOOC. In: Proceedings of the First ACM Conference on Learning@Scale Conference1, pp. 3–10. ACM (2014)

The World Bank Mobile phone access reaches three quarters of planet's population (2012). http://www.worldbank.org/en/news/press-release/2012/07/17/mobile-phone-access-reaches-three-quarters-planets-population (accessed April 20, 2015)

New Component Technologies and Development Strategies of e-Learning in MOOC and Post-MOOC Eras

Tsuneo Yamada

Abstract Information and Communication Technologies (ICTs) have triggered the innovations of pedagogies and learning methods in all levels of education. In addition to distance education, e-Learning is expected to improve classroom teaching through educational tools and digital content in various blended approaches; Mobile devices and SNS showed the new content distribution and knowledge sharing in learner communities; MOOCs (Massive Open Online Courses) expanded the opportunities of quality education on a global level. This paper discusses the Japanese practices in which MOOCs acted as catalysts implementing component technologies and development strategies for e-Learning.

Keywords Information and Communication Technology (ICT) · e-Learning · MOOC · Online course · Mobile learning · Blended learning · SNS · Metadata · Repository

1 e-Learning

"e-Learning" (electronic learning) is a technology-enhanced learning (TEL) which depends specially on Information and Communication Technologies (ICTs). With the progresses of basic learning theories and technologies, "e-learning" had many varieties and was thus known by many different names, such as CAI (computer-assisted/computer-aided instruction), CBI/CBT (computer-based instruction/training), WBT (web-based training), multimedia learning, CMI (computer managed instruction), internet-based training (IBT), flexible learning, online education, CMC (computer-mediated communication), cyber-learning, personal learning environments, virtual learning environments (VLE), m-learning (mobile learning), u-learning (ubiquitous

T. Yamada(✉)
The Open University of Japan, Chiba, Japan
e-mail: tsyamada@ouj.ac.jp

T.T. Zin et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 388,
DOI: 10.1007/978-3-319-23207-2_39

learning), and digital education. While the ultimate goals of e-learning are educational reform and innovation in education and learning, the interim goals are various, including lifelong learning and open education (education for all, "anytime, anywhere and anybody), a learner-centered approach, multiculturalism and pluralism, internationalization and globalization, personalization and optimization, and so on.

2 MOOC

The latest big wave of educational innovation was MOOC (Massive Open Online Course). After their first emergence in 2007, while MOOCs in the North America have passed a peak of inflated expectation (cf. A Gartner Report, "Hype Cycle for Education, 2013"), the regional MOOC Consortia are still launching in other regions, such as in Japan (JMOOC), Korea (KMOOC), China (CMOOC), Thailand (Thai MOOC) and Indonesia.

The main features of MOOCs are 1) Massiveness, 2) Openness, 3) Online services, and 4) Access to education, as evidenced in the name. Comparing with previous open educational resources (OERs), we can add several characteristics, such as 5) quality assurance as a course, 6) big data and learning analytics, 7) understanding of sustainability (acceptance of various business models), and 8) academic brand strength.

3 MOOCs in Japan

In 2013, MOOCs developed into a social trend in Japan. After the spring, Japanese universities started MOOC projects by joining the global consortia; the University of Tokyo released MOOCs from Coursera, and Kyoto University joined edX. Under the collaboration with the industries, they also launched the Japan Massive Open Online Course Consortium (JMOOC, http://www.jmooc.jp/en/) in November 2013 as a regional MOOC consortium. JMOOC is a "General Incorporated Association" in Japan, i.e., an NPO/NGO. By May 2015, 81 full members (39 academic, 5 public, and 37 corporate), 7 special contributing members, and 9 associate members had joined. JMOOC is maintained mainly by membership payments.

As of May 2015, JMOOC had three official platforms: "gacco," "OpeN Learning Japan," and "OUJ MOOC." "Gacco" (http://gacco.org/) is an Open edX-based platform managed by NTT DoCoMo and NTT Knowledge Square; "OpeN Learning Japan" (http://open.netlearning.co.jp/) is a domestic integrated learning support platform managed by Net Learning, Inc; "OUJ MOOC" is a multimedia e-textbook taste platform developed by CCC-TIES Consortium and managed by OUJ. JMOOC members can choose one of the official platforms in opening a MOOC as a course provider. The platform providers ought to offer the platforms to the members' requests at minimum cost. The JMOOC platforms do not have any mechanisms (such as APIs) to share content, tools, data, and so on.

4 OUJ-MOOC as a Pilot Platform

As a founding member of JMOOC, the Open University of Japan (OUJ) opened two MOOCs in the first releases from JMOOC (cf. Yamada, 2013c). As a unique open university in this country, OUJ has contributed to the Japanese open education and OER movements (Yamada & Yoshida, 2010; Yamada, 2013a). From the viewpoint of open education, OUJ has several basic questions about MOOC, such as: (1) "Can MOOC be a new sustainable model of open education?" (2) "Will MOOC show a new delivery model of higher/tertiary education to reach potential lifelong learners?" (3) "Can OER (in the narrow sense) and MOOC share the roles at open universities?" and (4) "Are open universities the providers or the competitors of MOOCs?" In order to examine the effects and influences, OUJ decided to launch the pilot MOOCs as a MOOC platform provider.

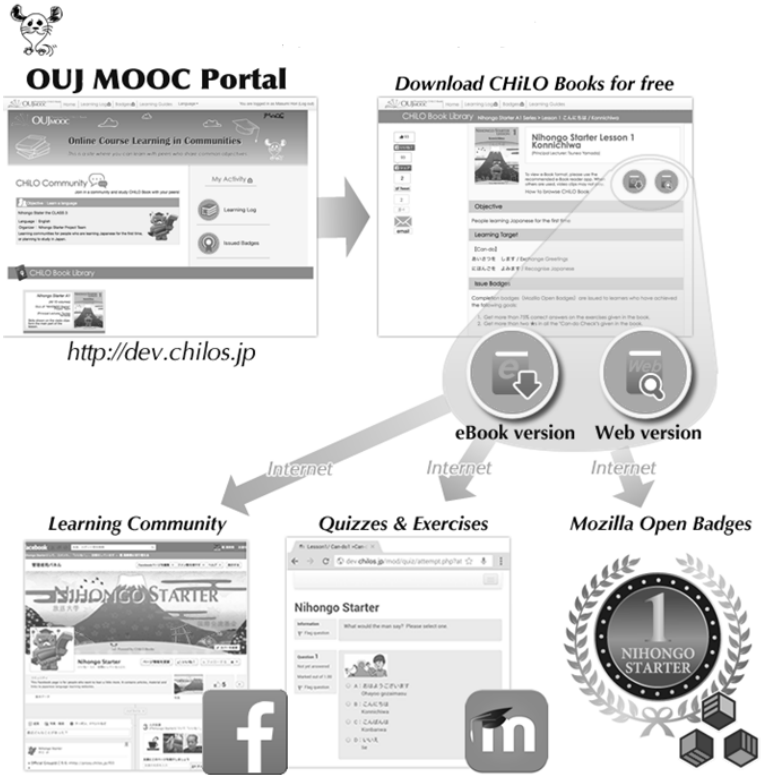


Fig. 1 OUJ MOOC platform powered by CCC-TIES "CHiLO Book" System: The concept (cf. Yamada, 2014).

The architecture of the OUJ MOOC platform is shown in Figure 1 (cf. Yamada, 2014). Considering the trade-off between the diversity of users' IT environments and cost, it was constructed by mash-up technologies of multimedia e-textbooks,

LMS (learning management system), and social networking services (SNS). Further, it consisted of iBook or e-pub 3.0 packaging (e-books), Facebook (registration and learner community), YouTube (video delivery), Moodle (LMS), and Mozilla Open Badge (certification). As some functions of e-textbooks can be used without connecting to the Internet, it was considered a better solution, especially in developing areas. We chose the "CHiLO Book" system (Hori, Ono, Kobayashi, & Yamaji, 2013; Hori, Ono, Kobayashi, Yamaji, Kita, & Yamada, 2014), which was developed by CCC-TIES. Some modules were added and/or deactivated among the courses.

OUJ launched two pilot courses from April 2014. NIHONGO Starter (A1) is an English course for non-native speakers of Japanese. International students who have no experiences of Japanese language learning can learn basic Japanese in English, which covers various topics and scenes that students may encounter when they stay in Japan. NIHONGO Starter (A1) is based on the JF Standard for Japanese-Language Education (http://jfstandard.jp/pdf/jfs2010_all_en.pdf). The standard was developed by the Japan Foundation (JF) and has common definitions for six levels of language proficiency with CEFR (Common European Framework of Reference for Languages). The MOOC is a short course of ten lessons and corresponds to the first part of Level A1 of the JF Standard for Japanese Language Education. Each lesson has two to four "Can-dos" (competences). The "Computer System" (Principal lecturer: Yoichi Okabe, President, OUJ) was developed using the course materials of his regular TV broadcasting course of the OUJ, and only a Japanese language version was available. Both courses were free of charge.

5 Evaluation

We had approximately 2500 registrants of NIHONGO Starter A1 from around the world. While the drop-out rates were still high (90-95%), the system managed learning processes and collected the data without apparent negative feedbacks from the registrants. Although we developed only an English version for NIHONGO Starter A1, the numbers registered from Central and South America, East and Middle Europe, and Arabic countries were more than those from English-speaking countries, and showed a new way of reaching potential learners who had been considered difficult to find out. Compared with the major surveys on the numbers of Japanese language learners overseas (e.g., The Japan Foundation, 2013), the registrants from China and Korea were much smaller. We considered that it depended on the kinds of social networking services and on language issues. When some digital device connected with the Internet, such as a personal computer, a tablet, and/or a smartphone, our learners could gain access to our digital learning resources and learn autonomously.

From the viewpoints of the business models and the sustainability, we had another issue. The numbers of the registrants were much smaller than we had expected and, as a result, our courses were not massive (hundreds to thousands). Most of the JMOOC courses in Japanese succeeded to have more than five or ten thousands of registrants. The results showed the difficulties of launching non-Japanese language courses in an original brand from Japan and disseminating them to the world.

6 Prospects for New Component Technologies and Development Strategies of e-Learning

6.1 The Goals Unachieved

When we had the commitments with MOOC phenomena, we predicted three development phases of MOOC and key technologies and social infrastructures in each phase (Yamada, 2013c). Table 1 was the revision of our previous works by considering the recent progresses. As of May 2015, technologies have achieved the goals of Stage 1: large-scale open education while the drop-out rates were still high. What is needed for Stage 2 is to improve the quality of learning processes and outcomes in sustainable ways. In order to optimize each learning process, that is, to realize effective personalization, we need innovation in key technologies, such as an AI engine for optimization, learning analytics tools, learning log (record) store for learning analytics, learning materials repositories, and the federation.

Table 1 The Growth Stages of MOOCs and the Impacts on Higher Education and Lifelong Learning (based on Yamada 2013c)

Stage	Impact on Higher/Tertiary Education	Indicator (e.g., Drop-out Rate)	Features, Key Technologies, and Social Infrastructures
1	A new type of OER *Still Limited*	80-90% and higher	Services of free education, not of open educational materials ● Large-scale LMS
2	A new model of open education *Sufficient Threat to OUs*	the same as the rates of correspondence courses and lower	Quality assurance of online courses and flipped classrooms; Customization of courses for improving learning processes ● Materials repository ● Learning analytics and learning log (record) store
3	A new model of higher education *Real Innovation*	the same as the rates of traditional universities, Face-to-Face/blended courses and lower	Personalization (Optimization) of lifelong learning based on truly learner-centered approach ● Federation of academic/educational databases (such as SIS, e-portfolio, and e-badge) ● Course accreditation by trusted third parties (micro-credentials, nano-degrees)

"Materials" Repositories and the Federation

In the near future, the personalization of learning will progress so that courseware can be customized to each learner's context. In order to personalize learning courses for each learner, course providers need to develop and collect sufficient components and modules, and must store them at a repository such as a "materials" repository. Under their limitation of financial and human resources, they have difficulty developing all the components and modules from scratch and need a common framework for sharing and reuse, also called an "eco-system." When the content is open and free, the simple federation of the "materials" repositories may be sufficient for the infrastructure; when it contains both open and proprietary content, the architecture needs some elements of e-commerce, such as encryption and billing.

In Japan, the National Institute of Multimedia Education (NIME, the predecessor of the Center for Open and Distance Education at OUJ), started an educational information portal service with content and metadata repository functions primarily for higher education in 2003 (Yamada, Yaginuma, & Inaba, 2003; Yamada et al., 2004). In March 2005, NIME launched a new gateway service on Japanese educational content called "NIME-glad" (Gateway to Learning for Ability Development, cf. Yoshii, Yamada, & Shimizu, 2008). These repositories stored various educational content, included learning and instructional materials, with metadata. NIME also started an IEEE-LOM based "JOCW Search" in October 2006 for the JOCW consortium. After the merger of NIME with OUJ on April 2009, under the collaboration with the National Institute of Informatics (NII), the backend system of "JOCW Search" was rebuilt with the WEKO repository system. WEKO is an open source repository system that was developed by the NII. The NII operates their nation-wide federation system of institutional repositories in Japanese HE institutions (JAIRO, http://ju.nii.ac.jp/en/) by using WEKO. While the previous usage of the repository system was mainly archiving the completed course materials, it can be used for "materials" repositories in the future. OUJ has started to operate a pilot repository for the research, and to discuss the improvement of metadata and metadata tagging, which can be used in automatic personalization processes (Yamada, 2013b).

Learning Log (Record) Store for Learning Analytics

The learning log (record) store is the data repository of learning logs or learning records of many learners' learning processes while e-portfolio is the database to store personal learning records and fruits. The logs are stored in multiple subsystems, such as learning management system (LMS), streaming servers, social networking services (SNSs), and other servers and proxies. In order to enhance the reusability, the logs are harvested and stored in the *learning log (record) store.*

In the beginning, in order to enclose the big number of learners, or "big data," some "killer" content was regarded as indispensable. At present, it is more difficult for many MOOC providers to use a huge budget for the development of the courses in the midst of severe competitive environments. They cannot sustain their MOOCs without the ecosystem. We have now recognized that "big data" can be realized by sharing learning log (record) by the federation of learning record

stores. The concept of "eco-system" will expand from content and tools to learning log data.

The studies on learning analytics have also been initiated, but in many course providers, the main concerns still remained as to how to measure learning activities ("sensor") or how to accumulate the data in reusable and sustainable fashions ("learning log (record) store").

6.2 Big Data and Privacy

The protection of personal data or privacy in big data collection is also an essential and pressing issue. The Japanese Parliament is currently deliberating the Amended Personal Information Protection Act as of mid-May 2015. OUJ and NII have exchanged a MOU to promote the innovations of academic teaching and learning, and to share learning log data through the common learning record store. Observing the progress of the deliberation, OUJ revised the personal data and privacy policy for sharing learning logs among institutions and for the future introduction of new student support services that are personalized by learning analytics (Yamada & Okabe, in preparation).

6.3 International Standards in e-Learning and Digital Publishing

Although as of May 2015 JMOOC had three official MOOC platforms, they had no clear interoperability with each other. In examining how to commit with international standardization activities on e-Learning and digital publishing, such as EDUPUB (http://idpf.org/edupub-2013-report) and IMS Global Learning Consortium (http://www.imsglobal.org/), we look for some collaborative frameworks to utilize the international standards for JMOOC's mission because our platform policy includes the concept of "joint." As the current concerns are learning metrics/analytics and API store, we have strong interests in IMS Caliper Analytics and experience API.

6.4 OER and Copyright Issues

Sharing the content at the component or module level is important to support localization or personalization of the course efficiently and effectively. In the reuse and remix, materials are often used in ways the original author never intended. Especially across borders, deployments in different cultural contexts can be pluralistic and unpredictable to the original creators. The propagation of OER-oriented course materials may depend on the tolerance of the original authors regarding the uncontrollability of their products. One of the ways to support the further propagation of OER, therefore, is to respect the creator's rights while supporting the opening of knowledge for the benefit of human society (cf. Yamada, 2013b).

Acknowledgements This study was partially supported by a Grant-in-Aid for Scientific Research (A) to the first author (Grant No. 26242013). The MOOC "NIHONGO STARTER A1" was developed through collaboration with the NPO CCC-TIES Consortium, the Japan Foundation, and OUJ. A part of the study was based on collaborative researches and/or operations with GLOBE (Global Learning Object Brokered Exchange) partner organizations, Japan OpenCourseWare Consortium (JOCW), and Japan Massive Open Online Course (JMOOC).

References

1. Hori, M., Ono, S., Kobayashi, S., Yamaji, K.: Development of Open Education Platform Utilizing GakuNin and e-book. IPSJ SIG Technical Report, pp. 1–8 (2013)
2. Hori, M., Ono, S., Kobayashi, S., Yamaji, K., Kita, T., Yamada, T.: Learner autonomy through adoption of open educational resources (OER) using social network services and multi-media e-textbooks. In: Paper presented at the 28th AAOU Annual Conference. OUHK, Hong Kong, October 27–31, 2014
3. The Japan Foundation. Survey report on Japanese-language education abroad 2012 Excerpt (in Japanese) (2013)
4. Yamada, T., Yaginuma, Y., Inaba, R.: A challenge for sharing digital learning materials of teacher education. In: Proceedings of 2003 KAEIB International Symposium and Conference "Educational Media in Schools", Seoul, Korea, September 4–5, 2003, pp. 189–190 (2003)
5. Yamada, T., Miwa, M., Aoki, S., Kato, H., Kawafuchi, A., Kodama, H., Kondo, T., Ohta, Y., Shibasaki, J., Yaginuma, Y., Inaba, R.: Sharing and re-use of digital learning materials in japanese higher education: a NIME new project. In: World Conference on Educational Multimedia, Hypermedia and Telecommunications (Ed-Media 2004), Lugano, Switzerland, June 2004, pp. 2323–2330 (2004)
6. Yamada, T.: Open educational resources in japan. In: Dhanarajan, G., Porter, D. (eds.) Open Educational Resources: An Asian Perspective. The Commonwealth of Learning & OER Asia, pp. 85–105 (2013a). (electronic open book)
7. Yamada, T.: An open "materials" repository and global search system: preparing for diverse learners and a variety of learning processes. In: McGreal, R., Kinuthia, W., Marshal, S., McNamara, T. (eds.) Perspectives on Open and Distance Learning: Open Educational Resources: Innovation, Research and Practice. The Commonwealth of Learning & Athabasca University Press, pp.153–163 (2013b). (electronic open book)
8. Yamada, T.: MOOC and open education: discussions in japanese higher education. In: Proceedings of the 27th AAOU Annual Conference, Islamabad, Pakistan, October 1–3, 2013, pp. 8 (2013c)
9. Yamada, T.: Development of MOOCs in a japanese open university. In: Paper presented at the OCWC Global Conference, April 23–25, 2014, Ljubljana, Slovenia (2014). http://conference.oeconsortium.org/2014/wp-content/uploads/2014/02/Paper_72-OU-Japan.pdf
10. Yamada, T., Okabe, Y., Hori, M., Ono, S.: OUJ MOOC platform: features and outcomes. In: Paper presented at the 28th AAOU Annual Conference. OUHK, Hong Kong, October 27–31, 2014
11. Yamada, T., Yoshida, M. (eds.) White papers of six Asia-Europe Countries: 02 e-Learning for Lifelong Learning in Japan (2010). In: Kim, B. (ed.) e-ASEM White Paper: e-Learning for Lifelong Learning. Korea National Open University Press, pp. 105–232

Development and Deployment of the Open Access Repository and Its Application to the Open Educational Recourses

Kazutsuna Yamaji, Toshihiro Aoyama, Masako Furukawa and Tsuneo Yamada

Abstract Worldwide activities on open access have triggered many universities to operate institutional repositories (IRs). The National Institute of Informatics (NII) has led a Japanese IR project since 2014 and, developing homegrown repository software named WEKO as a module for the content management system NetCommons (NC). Concepts of WEKO are "High Functionality", "Easy" and "As you like". WEKO has almost all functionalities you need as a repository system, and these can be customized and operated by browser. In addition, not only the repository functions but also variety of add-on can be utilized for designing your own web page. More than 250 universities in Japan are now operating WEKO as their IR. Since the WEKO has multilingual functionality, some of the Malaysian university has decided to employ it. In this paper, we summarize the repository related activity in Japan and point out the possible collaboration between open educational and repository.

Keywords Open access · Open science · Institutional repository · Cloud computing · Open educational resources

1 Introduction

As in the United States and in European countries, the institutional repository (IR) of university in Japan were created with high expectations of large-scale open

K. Yamaji(✉) · M. Furukawa
National Institute of Informatics, Tokyo, Japan
e-mail: {yamaji,furukawa}@nii.ac.jp

T. Aoyama
Suzuka National College of Technology, Suzuka, Japan
e-mail: aoyama@info.suzuka-ct.ac.jp

T. Yamada
Open University Japan, Tokyo, Japan
e-mail: tsyamada@ouj.ac.jp

T.T. Zin et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 388,
DOI: 10.1007/978-3-319-23207-2_40

access (OA). Our own Japanese institution, named the National Institute of Informatics (NII), has two core missions. One is to serve as a research center for information science; the other is to foster inter-university collaboration, which historically has consisted primarily of supporting information and communication technology infrastructure for higher education in Japan. Supporting IR activities of university libraries has long been an important project for NII. NII started a Cyber-Science Infrastructure program in 2005 [1] . Under the auspices of this program, we started a project which is directly funding institutional repositories in order to support community deployment and development [2] . The achievements of the project were remarkable. Through the end of the project in FY2013, the number of IR in Japan had grown to roughly 350, with combined registered contents of more than a million items. The majority of the content was departmental bulletins [3] , the typical publication media for the humanities and social sciences. Before this wide-scale IR deployment, most departmental bulletin papers were issued by means of paper based publication. The development of these repositories by university libraries brought reform to the humanities and social sciences by helping them transition from paper to electronic publication. The wide scale use of IR also increased the visibility of their work, transforming local publication to global publication. In addition to funding IR propagation, we developed a separate system named IRDB to aggregate metadata from all over Japan using the OAI-PMH (The Open Archives Initiative Protocol for Metadata Harvesting) protocol [4] and our custom metadata schema named Junii2 [5] . Our IR activity greatly increases contents availability in Japanese article search engine named CiNii [6] . The benefits to research in disciplines which had historically focused their publication in narrowly distributed departmental bulletin papers are obvious. Our IR project arguably triggered an evolution of scholarly content publishing workflow in these disciplines in Japan.

However, IRs developed by individual universities are entirely self-contained distribution systems. Each university must secure sufficient funds to launch and persist this service. There are more than 700 universities in Japan, which made our project fiscally unable to support every university despite the interest of each university's library staff. Our project succeeded in bringing an OA culture to universities through the implementation of IRs, but progress remained to be made, particularly with private universities. Anticipating the completion of the initial project and observing these lingering gaps, we began devising ways to further improve access to knowledge in Japan. In this paper, we introduce how to fill the gaps by utilizing our new IR cloud service named JAIRO Cloud using home grown repository software named WEKO, and our expansion to the open educational resources.

2 WEKO

2.1 System Architecture

Content management systems (CMS) such as Drupal, Joomla, and Plone enable users to construct web communities that make available many useful functionalities

to their users. In most CMS systems, website functionalities are embodied as modules separate from the core system and installed as the need arises. The NII has developed an AJAX-oriented CMS called NetCommons (NC) which is being used by educators. We developed a repository module for NC. The name "WEKO" comes from Swahili and means "repository" in that language. The system architecture of NC and WEKO are shown in Figure 1.

The system is written in a scripting language, PHP, rendering it OS-independent. MySQL is used as a relational database backend for storing data from NC and also WEKO. WEKO is open-source software under a New BSD (Berkeley Software Distribution) license. Installation merely requires a copy of WEKO to be placed in the NC modules directory, which becomes visible for activation in the administration menu.

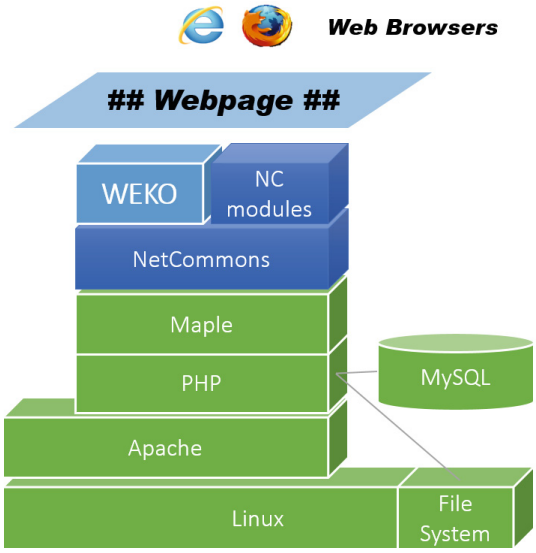


Fig. 1 System Architecture of WEKO and its Page View Example.

2.2 Functionality

An example of page view of WEKO repository module with other NC functions is shown in Figure 2. Administrators can build a web page using only a mouse because of its AJAX-oriented function. The administrator can compose the page design with different type of the functionalities which are prepared by NC. The IR is one of the primary faces of a university's scholarly communication, making the customization and design of the IR pages a crucial feature. Generally, to permit such customization, a service provider has to allow login-level access to a server to allow modification of template or HTML files directly. This has important

security implications for the service provider, so we sought to allow extensive customization instead just by use of an Internet browser. Examples of the top page design of the repository each of which is developed by WEKO and NC are shown in Figure 3.

Fig. 2 Example of Page View of WEKO Repository Module with other NetCommons Functions

Three different privileges of WEKO are summarized in Figure 4. One is the guest which can execute the directory search and keyword search including the full text search, and also can access to the ranking information. Full text searches work on PDF and MS Office application formats. Registered User has a privilege to register the item (contents). Status of the registered item can be seen in the workflow functionality. Metadata auto-fill and workflow functions are provided to support self-archiving by researchers. The administrator can access to the control panel of WEKO module. As has been mentioned before, the administrator can customize all of the WEKO functionality through this control panel. Most of the customization can be done on the web interface, allowing the administrator to operate the repository easily. There is no need to SSH into the web server. The administration menu includes item type (modify metadata set and OAI-PMH, edit tree (drag&drop tree edit and modify submission authority), content review, import, log analysis (csv and graphical output) and general settings (ranking calculation, log analysis black list,

site license list, full text library setting and so on). A site's design can also be customized without changing the source code. In order to offer even more functionality, WEKO can accept content deposited by means of the SWORD 2.0 protocol [7] and provide OAI-ORE resource maps [8] along with the index tree.

Fig. 3 Examples of the top page design of the different repositories each of which is developed by WEKO and NetCommons.

Fig. 4 Summary of the WEKO functionality of three different privileges.

3 JAIRO Cloud

An opportunity to increase OA to knowledge comes with the tide of cloud computing. Despite the progress made through the funding of multiple disparate IRs, the value of a centralized IR cloud service was obvious. Through our prior work and conversation with universities, we already knew that many universities desired to run an IR but didn't have the resources to do so. In response, we launched the JAIRO Cloud concept in 2010, a SaaS (software as a service) type IR cloud service [9] . Figure 3 shows service architecture of JAIRO Cloud. The actual deployment was accomplished using previously explained repository software WEKO. In 2011, we started pilot operation with several early adopter universities. Beyond providing only infrastructure, we offered workshops nationwide to assist end users in learning how to use the system. The community built through these workshops, with more than 200 total participants, used this knowledge base to in

turn train other local users. In 2012, we established a formal workflow and entered stable system operation, allowing the JAIRO Cloud to become a production-level service. The orange bars in Figure 6 depict the growth of the JAIRO Cloud. Remarkably and in contrast with the prior IR project, most of the participants in the JAIRO Cloud are private universities. This demonstrates our success in widening the spectrum of IR deployment in Japan.

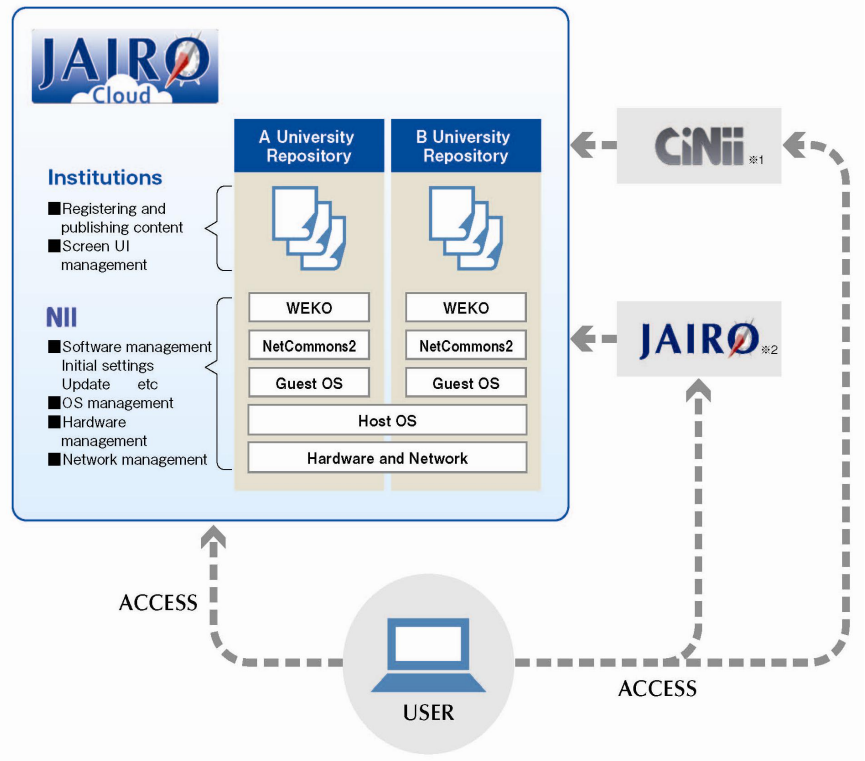


Fig. 5 Service Architecture of JAIRO Cloud.

OpenDOAR service statistics [10] show that as of the 24th of May in 2015, there were 2424 IRs deployed globally. As measured by OpenDOAR, the United States has the most such deployments, with 351 IRs. However, counting separately the multiple tenants of the JAIRO Cloud service, the number of Japanese IRs at the corresponding date was 498. As many of the IRs in Japan have not yet been registered with OpenDOAR, Japan could soon catch up with the United States in terms of IR activity. JAIRO Cloud participation has continued to increase very

rapidly despite the short time it's been in production. The accumulation and subsequent distribution of content has demonstrably improved the ability of researchers to search for and further develop knowledge. The JAIRO Cloud, although growing, is already a powerful tool in support of research.

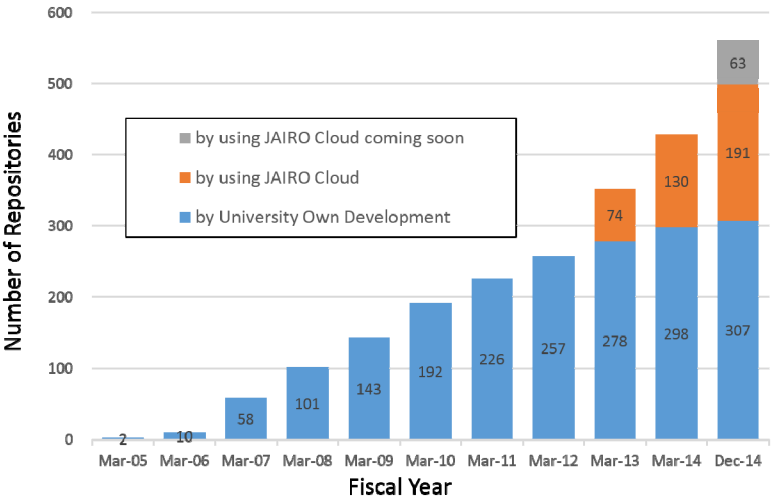


Fig. 6 Number of the Institutional Repository in Japan.

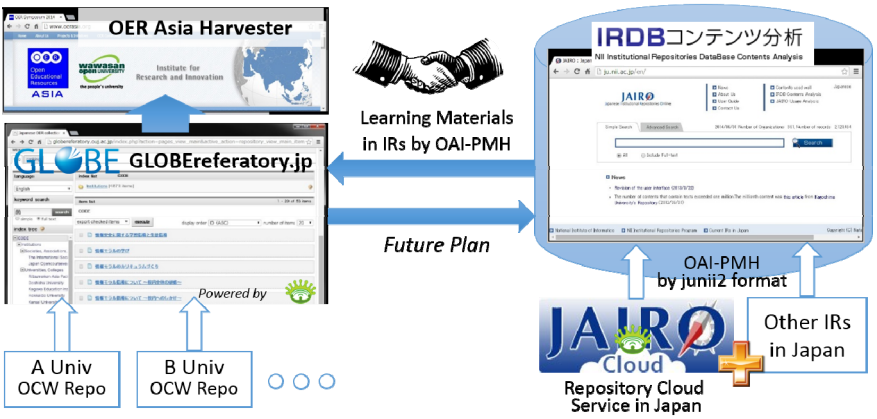


Fig. 7 Schematic diagram of this handshake model between GLOBEreferatory@Japan and IRDB.

4 Institutional Repository and OER Based Repository

The metadata schema defined by NII named junii2 has contents type definition such as journal article, dataset, learning material and so on. According to the IRDB statistical analysis, total ratio of the learning material is only 0.3%, however, its usage statistics is always high compared with that of the other item types. In order to increase the visibility of the learning materials in IRs, we started to collaborate with the OER (Open Educational Resources) repository.

As one of the GLOBE (Global Learning Objects Brokered Exchange) activity, Open University of Japan is operating the Japanese node repository (GLOBEreferatory@Japan [11]) which harvests OER metadata from OCW member universities in Japan. GLOBEreferatory@Japan also harvests OER contents from IRs via IRDB by using OAI-PMH, and enrich metadata in terms of the learning object metadata (LOM) schema. GLOBEreferatory@Japan employed our repository system named WEKO which can be the data provider and the service provider of OAI-PMH in both junii2 and LOM formats. Metadata in GLOBEreferatory@Japan is aggregated by OER Asia Harvester. Schematic diagram of this handshake model is shown in Figure 6. We are planning to feedback the OCW based metadata from GLOBEreferatory@Japan to IRs.

5 Conclusion

In this paper, we introduced our new repository system named WEKO and its application to the JAIRO Cloud service. Although the WEKO has been started to enhance the IR activity in Japan, its multilingual functionalities not only in English and Japanese but also in Bahasa Melayu, Cantonese, Chinese, Hindi, Indonesia, Tagalog, Thai, Vietnamese began accepted by Asian countries. We had several workshop in Malaysia and possible applicants start to install and operate the WEKO system as their repository system. We would like to enhance these collaboration and establish user community in Asia region.

The IR functionates as a show case of the institutional activity, therefore, it possibly involves different types of scholarly contents. On the other hand, researcher and/or educator prefer to utilize subject based repository rather than institutional based repository because of its visibility to their community. Case study of the handshake model between the two types of repository in this study will bring a new scholarly contents sharing eco-system. Since several Asian countries are interested in our handshake model, we will also propagate our collaboration in order to circulate and distribute contents metadata more efficiently.

References

1. Promoting the Cyber Science Infrastructure. http://www.nii.ac.jp/en/service/general/
2. NII Institutional Repositories Program. https://www.nii.ac.jp/irp/en/

3. NII Institutional Repositories DataBase Contents Analysis. http://irdb.nii.ac.jp/analysis/index_e.php
4. The Open Archives Initiative Protocol for Metadata Harvesting. http://www.openarchives.org/OAI/openarchivesprotocol.html
5. Junii2 metadata schema. https://www.nii.ac.jp/irp/en/archive/pdf/junii2_en_20090213.pdf
6. CiNii: Scholarly and Academic Information Navigator. http://ci.nii.ac.jp/
7. Allinson, J., Francois, S., Lewis, S.: SWORD: Simple Web-service Offering Repository Deposit, ARIADNE, vol. 54 (2008)
8. Open Archives Initiative: OAI-ORE. http://www.openarchives.org/ore/
9. Shiozaki, R., Tanabe, M., Mori, I., Yamaji, K.: JAIRO cloud: national infrastructure for institutional repositories in Japan. In: The 7th International Conference on Open Repositories (2012)
10. The Directory of Open Access Repositories. http://www.opendoar.org/
11. GLOBEreferatory@Japan. http://www.globereferatory.jp/

Challenges of Implementing e-Learning in Developing Countries: A Review

Than Nwe Aung and Soe Soe Khaing

Abstract The rapid developments of internet and communication technologies have materially altered many characteristics and concepts of the learning environment. E-learning has started to make way into developing countries and is believed to have huge potential for governments struggling to meet a growing demand for education while facing shortage of expert teachers, shortage of update text books and limited teaching materials. However, there are many challenges to implement e-learning in developing countries such as poor network infrastructure, lack of ICT knowledge, weakness of content development, etc. The objective of this study is to determine the major challenges of implementing e-learning systems in developing countries. The results of this study will serve as a basic for improving higher education in developing countries.

Keywords Developing countries · Higher education · e-Learning

1 Introduction

The developing countries are lagging behind developed countries in educational attainment and other aspects of the human capital development required in knowledge based global economy. A recent trend observed in higher education is the introduction of e-learning systems to provide students with online access to learning contents. The major driving forces behind this trend are the changing demographic factors of the students, changing conditions for education delivery and the innovation in technology itself. E-learning is seen as a tool for raising the number

T.N. Aung(✉) · S.S. Khaing
University of Computer Studies, Mandalay, Myanmar
e-mail: {mdytna,khaingss}@gmail.com

T.N. Aung · S.S. Khaing
University of Technology (Yatanarpon Cyber City), Yatanarpon Cyber City, Myanmar

T.T. Zin et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 388,
DOI: 10.1007/978-3-319-23207-2_41

of students who have access to higher education, especially marginalized groups in rural area s, by being a cheaper and more flexible alternative. Challenges are however plentiful; in many developing countries there is a lack of vital e-learning components such as computers, electricity and skills and the active participation of students and teachers. Universities in developing countries face unique challenges compared to developed countries and suffer from congested classrooms, e-learning can compensate the weakness of their traditional education methods and enables higher-education instructors to transfer their knowledge for a relatively large number of students without limitation of space, time or facilities. This study presents a review of research on challenges for implementing e-learning with a particular focus on developing countries. It is hope that the findings of this study will serve as a basic for educational institutions seeking out cost effectiveness alternatives to implement e-learning in developing countries.

2 Literature Review

In today's technology-driven age, e-learning has become an important tool for enhancing the delivery, interaction, and facilitation of both teaching and learning processes. The proper utilization of e-learning can promote time and location in-dependent access to the sources and contents of learning materials, reduce cost, and improve the quality of education (Cruthers, 2008).

However, despite its widely recognized advantages, implementing an e-learning project is not as simple as it seems. If not done properly, it can bring about a lot of problems and challenges the expected benefits can even fail to take effect (Graham, 2006). The biggest challenge probably lies in ensuring that certain preconditions are met for e-learning, such as access to ICT tools and network in-frastructure. Changing the perception of teachers and learners towards e-learning and convincing them to accept it is also very crucial. Another aspect that needs to be looked at is the technical competency of the people that will interact with the e-learning system (Gold et al., 2001).

This study proposed a conceptual framework of emerging issues for e-learning in developed and developing countries. The framework is useful to guide both practice and research. This research found 30 specific challenges which were grouped into four categories: courses, individuals, technology and context. The overall conclusion is that these challenges are equally valid for both developed and developing countries; however in developing countries more papers focus on access to technology and context whereas in developed countries more papers concern individuals (Anderson & Gronlund, 2009).

The report presented a collation of informational resources that document the potential of e-learning in developing countries, factors affecting its impact on education, the economy and society, and what experience has shown to lead to a successful integration of e-learning into educational systems(Olson et al, 2011). This study observed in developing countries like Pakistan, which have not yet been able to benefit fully from the advantages of e-learning. This study identified

the issues, related to e-learning through the feedback captured from students and provided strategies to successfully overcome the issues (KholaIlyas, 2013). This study identified the major challenges facing development and adoption of e-learning in Private Universities in Kenya: ICT infrastructure, e-learning curriculum, Instructors' competencies, Performance expectancy, perceived usefulness of e-learning by students (SorillaNisperos, 2014).

This study presented many teaching methodologies, which are using in Engineering Program of Cambodia's Higher Education, then showed the comparison result of student learning output, also presented the lecturer's and professor's mindset on moving from traditional to flipped methodology, and finally they proposed an appropriate teaching methodology(HEAN and SOK, 2014). This study is started from how many percentages of Higher education institution student using computer and internet access penetration (Chanthamalay, 2014). This study carried out research and development framework of the e-learning course distribution of the lessons learned from two years' experience and identified the current challenges and to make some suggestions for the future (Khaing and Win, 2014). This study presented the description of a survey instrument that has been developed to assess e-learning readiness with the cooperation of students from different faculties in Hanoi University of Science and Technology (henceforth HUST). The study showed that HUST's student overall ready for e-learning but they need to enhance their ability in adapting with new way of learning instead of conventional learning (Ngo et al, 2014).

3 Challenges Facing Implementing e-Learning in Universities

Literature from the previous section has identified the following major challenges facing deployment and adoption of E-learning in universities of developing countries.

- **ICT Infrastructure**
  The cost of acquiring, managing and maintaining ICT Infrastructure has been identified as the major stumbling block in deployment and adoption of e-Learning by institutions of Learning. Some of the factors such as poorly equipped classrooms and lack of electricity have hindered the deployment and subsequent adoption of e-learning especially in rural areas.
- **Course**
  The first issue identified here is the curriculum which stipulates much of the course actions and content. Many curriculum developers are using the dame models to create e-learning instruction as they used to design and develop face to face teacher and learner instruction. Lack of a proper e-learning curriculum is a major barrier to effective deployment and adoption of e-learning. Another course issue is the delivery mode of the course. The factor concerns whether students should be allowed to choose the

medium of content delivery. This factor is often discussed in a context of the global mobility of learners where the education is not nation-bound.

- **Individuals' Characteristics**
  The characteristics of the individual student, and in some cases the teacher, much researched in developed countries, less so in developing ones. Student motivation is a factor that is frequently discussed in surveys on what affects students' satisfaction and capacity. The relation between motivation and other e-learning factors is rarely elaborated; the reasons for success or failure in the studies are simply referred to as "personal motivation" or "lack of motivation".

- **Contextual Factors**
  The context of e-learning includes the context of the delivering organization as well as the context of the society in which the e-learning takes pace, including culture, traditions, rules and regulations. Research addressing the delivering organization is mainly concerned with the organization and management of the delivery side's functions and the need for changes in organizational structures. A frequently addressed issue here is that of the organization's knowledge management or knowledge building. This factor is addressed in terms of the need for a knowledge repository built on research and evaluations and to establish e-learning units.

- **Instructors' Competencies**
  Teachers who are insufficiently trained in their own subjects, and have little or no computer experience are a hindrance to effective deployment and adoption of e-learning in institutions of learning. The integration of ICT technologies would require significant pre-service and in-service teacher training in basic computer literacy as well as how to teach with e-learning technologies for effective deployment and adoption of e-learning in universities.

- **Technical Skills**
  Technical skills are a significant aspect of implementation and integration of e-learning technologies in education system. They include installation, availability of latest technology, fast internet connection, uninterrupted supply of electricity, maintenance, administration, security and absence of technical support. Most of the developing countries lack quality experts for implementations and maintenance of Information and Communication Technologies (ICT).

- **IT Literacy**
  The degree of proficiency in computer technology is an important factor in successful adoption of technology. The confidence in skills and ability to use e-learning will contribute significantly towards the usage of technology. Most likely the more experience the users have in using the Internet and computer, the more likely they will accept and use e-learning.

- **Language Competency**
  Students having low proficiency are not likely to use e-learning because of low confidence in understanding the contents of English written materials. The study found that most of the respondents felt language was a barrier to e-learning. This finding is consistent with studies in other developing countries. For example, the UNESCO report indicated a need for adequate Thai courseware for e-learning in Thailand.

- **Awareness**
  Knowledge and understanding of the e-learning benefits motivate the students to participate. Students unaware of the benefits of e-learning are likely to get frustrated easily as they may take it as a time wasting activity. Without realizing the importance of a particular technology and its contribution to the achievement of goals, successful integration of technology is difficult.

- **e-Readiness**
  The level of readiness of higher education institutions were divided into four dimensions, namely: the perceived e-readiness of faculty and students, their level of acceptance of the technology, the need for training and the readiness of the technological infrastructure of the university to support e-leaning. To achieve a higher level of readiness, universities need to provide preparatory training to both faculty and students to further improve their skills in handling the technology involved in this environment.

4 Findings and Recommendations

Based on this study, the above challenges are facing deployment and adoption of e-learning in developing countries. It is crucial for educational institutions and governments to address the above mentioned issues in the most effective manner for the specific country contexts. The key question is how e-learning approaches can help address these challenges, and provide students a leap forward in their universities leaning and in their future employment opportunities. The e-learning approaches need to be designed to fit the local situation and needs, for example content needs to be not only contribute to the curriculum and in the local language but it also need to reflect cultural norms. Strong teacher training and professional development, mentorship, networking and support to integrate e-learning pedagogical approaches into classroom practice and curriculum are successful activities of the teacher. Many studies of e-learning programs have concluded that the key to ensuring successful outcomes is to blend more traditional classroom approaches with those that use technology.

5 Conclusion

Government and donors in developing countries realize the critical importance of education for economic and social development. The benefits of e-learning are believed to be great enough to allow the governments of developing countries to meet the growing need of education effectively. Educational institutions and governments need to coordinate their efforts to address the existing issues in order to promote and support e-learning initiatives. The brief history of e-learning programs in developing countries has provided some lessons in what activities work, and what produce sustainable programs. Sustainable e-learning programs involve strong national leadership and many participating actors. National institutions include teacher training, the Ministry of Education and the private sector. International partners can play a vital role providing technical expertise and financial support.

References

1. Andersson, A., Gronlund, A.: A Conceptual Framework for E-learning in Developing Countries: A Critical Review of Research Challenges. The Electronic Journal on Information Systems in Developing Countries, EJISDC **38**(8), 1–16 (2009)
2. Newton, D., Ellis, A.: Effective implementation of e-learning: a case study of the Australian Army. Journal of Workplace Learning, col. **17**(5/6), 385–397
3. Baggott, G.: Biostatistics: e-learning strategies for improving student understanding – an e-learning practice case study
4. Tint, H.: Present Situation of Distance Education in Myanmar
5. Qureshi, I.A., Ilyas, K., Yasmin, R., Whitty, M.: Challenges of implementing e-learning in a Pakistani University. Knowledge Management & E-Learning: An International Journal **4**(3), 310–324
6. Olson, J., Codde, J., deMaagd, K., Tarkleson, E., Sinclair, J., Yook, S., Egidio, R.: An Analysis of e-Learning Impacts & Best Practices in Developing Countries
7. SorillaNisperos, Lea: Assessing the E-Learning Readiness of Selected Sudanese Universities. Asian Journal of Management Sciences & Education **3**(4), 45–59 (2014)
8. Phuoc, N.M., Hung, N.H., Anh, N.K., Tan, P.X.: Student's readiness on e-Learning in techincal university (Special at HUST). In: International Conference of Educational Technology ICET 2014, p. 99 (2014)
9. Namisiko, P., Munialo, C., Nyuongesa, S.: Towards an Optimization Framework for E-Learning in Developing Countries: A Case of Private Universities in Kenya. Journal of Computer Science and Information Technology **2**(2), 131–148 (2014). ISSN: 2334-2366 (Online)
10. Chanthamaly, P.: Study if the situation of higher education institution students accessing to internet for e-learning system development in lao PDR. In: International Conference of Educational Technology ICET 2014, p. 73 (2014)
11. Stanley, R., Lynch-Caris, T.: An innovative method to apply the flipped learning approach in engineering courses via web based tools. In: Proceedings of the 2014 ASEE Gulf-Southwest Conference, pp. 1–9

12. Samboeun, H., Kimheng, S.: Using flipped methodology for teaching and learning in engineering program: case study in ITC, Cambodia. In: International Conference of Educational Technology ICET 2014, p. 97 (2014)
13. Khaing, S.S., Win, A.: A research & development framework of successful rolling our strategies for e-learning courses. In: International Conference of Educational Technology ICET 2014, p. 79 (2014)
14. Ahmed, T.T.: Toward Successful E-Learning Implementation in Developing Countries: A Proposed Model for Predicting and Enhancing Higher Education Instructors' Participation. International Journal of Academic Research in Business and Social Science **3**(1), 422–435 (2013). ISSN: 2222-6990
15. Survey Research on e-Learning in Asian Countries – Fiscal Year 2002 (Country Specific Report – Myanmar)

SWOT Analysis of E-Learning Course Operation in Higher Education (Case Study: University of Technology, Yatanarpon Cyber City)

Soe Soe Khaing, Aung Win and Than Nwe Aung

Abstract E-Learning is the fast and essential method of delivering educational contents. E-Learning supports one of the alternative ways of traditional teaching and learning. So, most developing countries are initiated e-learning system. This research work aims to provide strengths-weakness- opportunity-threats (SWOT) analysis, the reflection of students and teachers, current infrastructure of e-Learning course operation in higher education in Myanmar. This research is based on online course operation run by three year experiences of ASEAN-Korea Cyber University (ACU) project at the University of Technology (Yatanarpon Cyber City), Myanmar. It discusses the details of project implementation and its capacity to support a new pedagogical framework ('before', 'during', and 'after') course operation at this university. It also concludes with a set of proposed recommendations for the future. Some important issues have been answered and evaluated.

1 Introduction

E-Learning service plays an important role in providing to enhance lifelong learning and provide unlimited opportunities for personal growth and development to all. [7]

University of Technology (UT) (Yatanarpon Cyber City) is one of the Member Universities of ASEAN-Korea Cyber University Project (hereafter ACU project). E-learning center at UT was established on July 11, 2012. Our motto is "Brighten your future with cyber education". Based on this motto, it can facilitate the students for learning interaction (learner participation) in the operating (learning) process.

S.S. Khaing(✉) · A. Win · T.N. Aung
University of Technology (Yatanarpon Cyber City), Mandalay, Myanmar
e-mail: {khaingss,yeyint2,mdytna}@gmail.com

S.S. Khaing · A. Win · T.N. Aung
University of Computer Studies, Mandalay, Myanmar

T.T. Zin et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 388,
DOI: 10.1007/978-3-319-23207-2_42

1.1 Background

At the 12th ASEAN-ROK Summit, the leaders of ASEAN agreed to continue to explore the possibility of establishing a cyber-university in order to promote international capabilities of ASEAN nations and empower Asian countries to play active roles in knowledge-based societies. The project plan was set with the three phases: the first phase is from 2010-2012.9 to carry out the basic research related with infrastructure building. It's the pilot phase of the project. The second phase is from 2012-2014 to establish the ASEAN Cyber University. The third phase is from 2015~2019 to make an expansion and growth.

1.2 Objectives

- Co-operate relationship between the Cambodia, Laos, Myanmar, Vietnam and Korea.
- Strengthen the education capability based on ICT (Information Communication Technology) in Myanmar.
- Establish e-Learning center.

1.3 Vision

The vision of the project is achieving shared growth involving ASEAN-Korea through academic collaboration.

1.4 Establishment of the Project

To establish e-learning center at UT, Korea International Cooperation Agency (KOICA) provided studio room, Learning Management System (LMS) system, other facilities for e-learning operation and the necessary training for the staff in

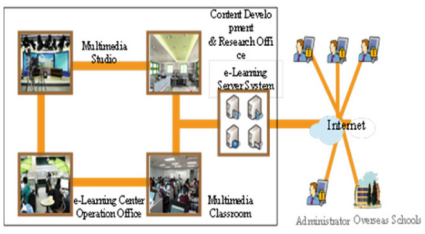


Fig. 1 Design and Architecture of ACU project e-Learning Operation

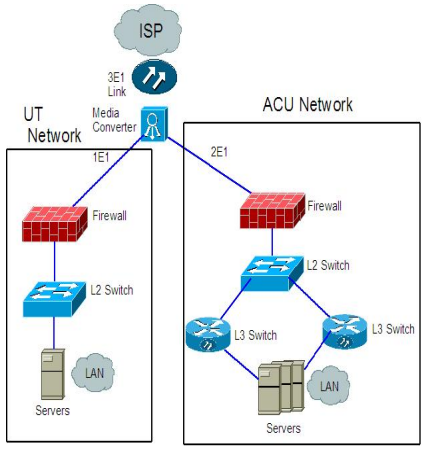


Fig. 2 Network Infrastructure at UT

e-Learning operation. Firstly, they dispatched e-learning center experts to Myanmar providing training programs. The training programs were divided into three parts, for managers, operating engineers, instructional designers and content developers.

The design and architecture of e-learning operation for ACU project can be seen in the following Fig. 1.

1.5 Host University of ACU in Myanmar

E-Learning implementation was initiated in 2012-2013 academic year. UT operated online courses in two semesters per academic year. The first semester begins in January and ends in mid-May. The second semester begins in May and ends in October. This research mainly focuses on sharing the three year experiences of e-learning operation at UT.

The objectives of this research are based on the followings.

- Investigate how UT e-learning operation is running.
- Highlight the SWOT Analysis of operation.
- Get the feedback of the students and teachers.
- Show the analysis of current operation.

2 Current Infrastructure and Facilities

2.1 Network Infrastructure at UT

E-Learning course operation at UT is run on the network infrastructure with 3E1 lines, 6Mbps link from Myanma Post and Telecommunication (MPT) Internet Service Provider (ISP). It separates 1E1 link, 2Mbps connection to the local campus and 2E1 links, 4Mbps connections to ACU project. Current network infrastructure of UT can be seen in Fig. 2.

2.2 Computer Laboratories

There are two computer laboratories in which each includes 85PCs with network access. The specification of each PC is Core 2Duo 2.4GHz, 4GB of memory and 500GB of Hard Disk space with Windows 7 Operating System. UT allocated one pc for each student by using shifting system. The students can use computer 9 hours per week and more than 2 hours for e-learning. Because of the international bandwidth limitation and internet speed, firstly UT utilized local LMS that was supported by KOICA for 4 semesters. But because of so many bugs in local LMS, ACU developed new LMS and currently new ACU LMS is being utilized.

3 Development Methodology

The SWOT analysis carried out in UT e-learning center to help the decision makers on the areas of strength and greatest opportunities with respect to e-learning. The analysis was based on the experiences with the university in general as well as the perspectives of the instructors, students, administrators and technical staff towards using online and offline survey. The outcomes of the analysis are illustrated in the followings.

Strengths

- Have an e-learning infrastructure is in place
- E-learning is a more flexible way of delivering contents especially frequently access contents
- Involved within ACU strategic plan and project initiatives
- Have in-house content developers (recently)
- Piloted and have an understanding of the ACU learning management system
- The staffs involved in delivering e-learning have a lot of experience in the field i.e. 3 years running in e-learning operation
- Have strong motivation and dedication
- Improve students ability to face challenges
- Provide student willingness to accept changes
- Get the availability of computer based infrastructure

Weaknesses

- Limited international bandwidth and internet speed
- No fulltime staff in e-learning
- The teachers have lack of experience in teaching with new technologies.
- The teachers have lack of willingness to learn and adopt new teaching methods (flexible learning)
- They don't have insufficient space and time for learning new technologies.
- The language barrier for the teachers and the students
- The lack of motivation for the students
- The lack of tutors (graduate students) with experience in e-learning
- Have inadequate in critical or analytical thinking
- Lack of quality e-learning training material

Opportunities

- Experience in faculty development opportunities (professional meeting, attending professional development training, seminars and workshops)
- Have funding opportunities for improving the educational methodology
- Experience of using Learning Management System
- Can engage subject matter experts within university who then can use e-learning as a blended solution
- Increase volumes of those receiving training

- Ability to link competencies to training
- Competent faculty for developing content
- Improvements in students' learning
- Increased student enrolment in the university

Threats

- Future budgetary limitations
- E-learning being seen as not cost effective.
- E-learning seen as the second best for delivering knowledge or skills
- Changes in policy and regulation in higher education

The SWOT analysis indicated that the current e-learning operation was likely to be successful, considering the faculty's strengths: wide use of computers for teaching and learning, well-equipped computer laboratories and, good computing skills among instructors and students. However, it is needed to take action to minimize its weaknesses including language barrier among students and instructors and lack of experience in teaching and learning with new technologies. The success of e-learning initiatives was also likely to be improved by exploration of the available opportunities and using them to counter the potential threats. An introduction of a reward scheme for instructors willing to develop e-learning expertise might be one opportunity worth pursuing; such a scheme might counter the threat of losing existing instructors and encouraging others to be involved in e-learning initiatives.

3.1 Analysis of the Survey Result

We conducted the survey to get the reflection of e-learning operation at the end of the first and the second semester. There are 67 students responded to the online survey over a two week period and 206 students responded to the offline survey. Based on the analysis of the survey result almost all the students are the age of less than 20. We conducted the survey on 67.4% of female students and 32.6% of male students. They responded one third percent of them are using computer from home and university respectively. Almost 30% of the students are working with computer on the average of 1-2 hours per day. Over 40% of the students responded the use of computer everyday and they are online 1-2 hours per day.

3.2 IT Literacy Survey Result

According to the survey result, over 62% of the students are good at using search engine, word processing and 40% of them are familiar with using computer. Most of the students are competent in IT literacy.

3.3 E-Learning Satisfaction Survey Result

60% of the students agree using e-learning system is easy. Over 50% of the students agree the screen layout is easy, the instruction of the content is easy to follow and the contents of e-learning match their needs. Almost 50% of them agree this e-learning course provides clear instruction.

Table 1 Student Satisfaction Survey Result

statement	Percent Responded on agree
Getting started with this e-learning system is easy.	**60.07**
This e-learning material is engaging.(I'm spending quite a time with this e-learning material)	**39.19**
I think I'm learning with this e-learning system.	**61.54**
Navigating through the given menu is easy to do.	**50.18**
The contents of e-learning match my needs.	**55.31**
Finding the options that I want in the e-learning system is easy.	**56.78**
Screen layout (e.g. going to NEXT page, play control bar, speed control) is easy to use.	**52.4**
This e-learning course provides clear instructions (i.e. lesson goal, quiz or summary, main contents)	**49.45**
The instruction of the contents is easy to follow.	**55.68**

Moreover, 57% of the students satisfied and concentrated on learning in this e-learning system.

3.4 Suggestion for Future E-Learning

58% of the students responded 'yes' for the question "do you think e-learning is going to be expanded for the future". The students prefer to expand e-learning in the future.

When the students responded the question "what kind of difficulties should be improved to expand e-learning", 62% of them would like to have the interaction between learner and instructor.

When we analyze the result of the question "what kind of activities should be reinforced to improve the effectiveness of e-learning", 43% of students prefer both project based and simulation/practicing based learning.

3.5 Circumstances Survey

Over 55% of the students responded 1~ 5 times per week they use online and 53% of them answered their computers are fast and 90% of them have laptop computers. 46% of students used Free Wifi and 43% of them have less than 4 inches

smart phone. The circumstances survey result is shown in the following table. The table is shown only the evidence result responded from the students.

Table 2 Circumstances Survey Result

Statement	Satisfied	Unsatisfied
Introductory guide about the credit exchange of the ASEAN Cyber University	**74**	7.33
Availability of courses offered	**77.3**	11
Credit transfer from the other institutions	**64.1**	**18.32**
Ease of use for the LMS system for registration	**70**	**18.32**
Information about the courses offered	**73**	8.43
Guide for the learning contents interface (menus or buttons)	**68.5**	6.96
Quality of learning contents that support efficient and effective comprehension and memorization	**74**	11
Server stability (No buffering or errors) during learning contents	**54.6**	**39.6**
Ease of questioning and answering during learning contents	**72.53**	17.6
Ease of use for LMS that supports learning activities (discussions, Q/A, homework)	**68.13**	**24.2**
Guide for the discussion board (discussion topics, due date, participation tips)	**65.2**	**25.64**
Interaction with the instructor	**63.74**	**23.81**
Interaction with peer learners	**72.16**	**18.32**
Ease of use for the system functions (play control, speed control, page navigation buttons, table of contents, learning support tools)	**71.43**	11

According to the result, over 70% of the students satisfied on the introductory guide, interaction with peer learners, ease of use for the system functions, quality of learning contents and ease of questions & answers. But almost 40% of the students' unsatisfied on the instability of the LMS server.

3.6 Reflection of the Instructors

The overarching research questions to the 10 co-instructors of e-learning course operation that we have selected for the case study research are:

- Did you satisfy to manage e-learning?
- Did you get to know what the probable problems of last semester's e-learning were?
- Have you got some amount of knowledge enough to manage e-learning?
- Have you got some ideas about how to operate your e-learning course next semester/year?
- Do you believe that you can operate better in the future than before?

Most instructors didn't satisfy providing e-learning operation because of no experience and limited time offer to the students. Actually they would like to provide more time to the students. Also they all have to support not only online class but also offline class. So they have limited time to support e-learning operation even though they are interested in e-learning. They need to have some strategy to achieve more involvement in e-learning.

4 Suggestions for Future E-Learning Operation

The suggestions have to be made the followings:

Learning Environment (Infrastructure) can be fulfilled with the providing of more bandwidth. Nowadays there is the competitiveness of telecom companies in Myanmar: Telenor, Ooredoo and MPT. Also the mobile density is getting increased and internet users are more and more increasing. They are providing incredible amount of internet speed and hence the students can access e-learning anytime, anywhere. To provide seamless access to LMS, the mobile version of e-learning contents need to develop.

To enhance students' engagement and motivation in e-learning, the instructors should make more discussions and assignments. And also students should get various incentives such as extra point or score for their active participation, interaction, and collaboration with peer students. Also more contents need to be developed locally and it can facilitate the students understand easily by the explanation of local language. The teachers will have to get more experience in developing e-learning with the help of development tools. To provide blended learning, the curriculum mapping between offline contents and online contents are required. The students will have more time in learning lectures. ASEAN Cyber University (ACU) provided new LMS this year so it would be able to overcome instability of LMS problems. The more course running in e-learning the more experience will be achieved by the staff.

5 Discussion and Conclusion

The reflections of the participating students and instructors on the e-learning pilot revealed that the initiative has achieved its objectives with respect to supporting framework. It is efficient online provision of course content particularly satisfied with requirements.

E-learning is one of the most important success factors for the economic development in Myanmar. It will help to improve the quality of education system and will also apply to vocational training for industrial workforce cultivation.

UT is the only one e-learning center under ministry of science and technology (MOST). Therefore, UT will be the main center of e-Learning contents distribution. UT will share online contents lecture to other universities. Moreover, UT would be a hub center of e-Learning among other universities. By doing this research, it will provide the strengths, weakness, opportunity, threat and success of e-learning operation in which what we should have to prepare, what we need more and how the problems to be fixed in the future. In spite of show-stopping issues, UT achieved very successful e-learning project. Without dedicated efforts from UT's faculty and supports from ACU, it's impossible to have 5 successful e-learning semesters. UT is also the frontrunner of e-learning implementation in CLMV countries and other ACU member institutes. Besides this, it is truly fruitful experience of UT's e-Learning contents operation for the first time and it will be a special chance to share this to other member institutes and ACU project.

References

1. Rhema, A., Miliszewska, I.: Reflections on a Trial Implementation of an e-Learning Solution in a Libyan University. Issues in Information Science & Information Technology **8** (2011)
2. Kim, J.: Final Report: ASEAN Cyber University e-Learning Project in UT YCC, Myanmar (December 2013)
3. Park, J.S., Kim, Y.S.: Policy & Curriculum Expert Dispatch Program I. Project for ASEAN Korea Cyber University (December 2011)
4. KOICA: The Strengthening CLMV capacity for Establishing of ASEAN-ROK Cyber University (Pilot Project) (June 2011)
5. Mohammad, S., Job, M.A.: Evaluation of Infrastructure for e-Learning System in AOU Bahrain Branch. International Journal of Information & Communication Technology Research **2**(4) (April 2012)
6. Mohammad, S., Awadhi, A.A.: Performance Measurement of Learning Management SYstem in AOU-Bahrain Branch. International Journal of Information and Communication Technology Research. **3**(1) (January 2013). ISSN 2223-4985
7. Khaing, S.S.: Current Status of Higher Education and e-Learning System in Myanmar. Senior Officials Meeting on ASEAN-Korea Cyber University Establishment, Seoul, Korea, January 17–21, 2011
8. Khaing, S.S.: Analysis of e-learning course operation for higher education. In: e-Learning Conference, Korea (2013)
9. Khaing, S.S., Win, A.: A research & development framework of successful rolling out strategies for e-learning courses. In: Proceeding of the First International Conference on Education Technology, ICET 2014, Seoul, Korea
10. UT: ASEAN-Korea Cyber University Survey Questionnaire, from AUN (July 2011)
11. Jonathan, W., Margaret, A., David, M.: Student e-Learning Survey Report (May 2004)

A Sematic Role Labeling Approach in Myanmar Text

May Thu Naing and Aye Thida

Abstract There is a generally certainty in the natural language and computational linguistics communities that semantic role labeling (SRL) is an important step toward improving significant applications, e.g. question answering, text summarization and information extraction. We propose a new method for assigning semantic roles on the structured trees of Myanmar sentences using Myanmar Verb Frame (MVF). In this paper, there is not use any machine learning techniques for SRL. It employs with predicate-argument identification algorithm and mapping algorithm to identify semantic roles in Myanmar. These algorithms mainly work on the syntax structure of Myanmar sentences. This system achieves over 70 % success rate in labeling the semantic role of pre-segmented constituents on the datasets.

Keywords Semantic roles · Myanmar Verb Frame · Predicate-argument

1 Introduction

The natural language processing community has recently experienced a growth of interest in semantic roles, since they describe WHO did WHAT to WHOM, WHEN, WHERE, WHY, HOW etc. for a given situation, and contribute to the construction of meaning. The semantic role represents the relationship between a predicate and an argument. It provides a general semantic interpretation of the sentence, and it can play a key role in natural language processing (NLP).

Semantic role labeling (SRL), sometimes also called shallow semantic parsing, is a task in natural language processing consisting of the detection of the semantic arguments associated with the predicate or verb of a sentence and assigning their

M.T. Naing(✉) · A. Thida
Natural Language Processing Research Lab,
University of Computer Studies, Mandalay, Myanmar
e-mail: {mtn.maythunaing27,ayethida.royal}@gmail.com

T.T. Zin et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 388,
DOI: 10.1007/978-3-319-23207-2_43

specific roles to the constituents of sentences. The Shared Tasks of CoNLL 2004 and CoNLL 2005 defined the task of SRL as "analyzing the propositions expressed by some target verbs of the sentence. In particular, for each target verb all the constituents in the sentence which fill a semantic role of the verb have to be recognized". SRL has widely used in applications like Question Answering Systems, Machine Based Language Translation Systems, Document Summarization and Information Extraction.

SRL approaches had been presented in other languages. But Myanmar language does not have any SRL system. And most of the SRL use classification approaches on Lexical Semantic Resource (LSR) such as FrameNet or Propbank. The proposed SRL approach does not use classification methods for assigning the semantic roles to Myanmar sentences. In this paper, we present an approach for SRL, in which shallow syntactic parsing and lexical resources are used.

This paper is structured as follows: In Section 2, we discuss the related works for SRL. Section 3 talks about Myanmar Verb Frame Resource. Section 4 describes the SRL method for Myanmar sentences. Section 5 outlines evaluation of the system on the test set. In Section 6, we conclude and talk about future works in this area.

2 Related Works

Most systems for automatic SRL at that time made use of a full syntactic parse of the sentence in order to define argument boundaries and to extract relevant information for training classifiers to disambiguate between role labels. Thus, the task has been usually approached as a two phase procedure consisting of recognition and labeling of arguments.

The SRL task was the shared task of CoNLL 2004 (Conference on Computational Natural Language Learning) and CoNLL 2005 conferences. The papers defined the SRL task and defined the types of semantic roles for a verb in a sentence. Several systems participated in the shared task and proposed several guidelines for the SRL task.

The CoNLL-2004 shared task discussed several models for the learning component including pure probabilistic models [1],[2]and [3], Maximum Entropy models[4], generative models [5], Decision Trees [6] and Support Vector Machines [7] and [8]. The method of calibrating features was defined by Nianwen [9]. The Co-NLL 2005 conference SRL provided an insight on several statistical methods followed, notably the Maximum Entropy classifier of Wanxiang [10]. The Co-NLL 2005 shared task introduced several techniques including Support Vector Machines (SVM), Conditional Random Fields (CRF), Perceptron Based Learning, etc. The best results were obtained by systems using Support Vector Machines.

The first method for SRL based on FrameNet was proposed by Gildea et al. [11]. They achieved argument identification and semantic role assignment with conditional probabilistic models. Their method includes example boosting in order

to cover the shortage of annotated examples. Kwon et al. employed ME methods for semantic role parsing based on FrameNet [12]. Pradhan et al. and Bejan et al. proposed semantic parsing based on FrameNet or PropBank [13] with SVM [14], [15] Hizuka et al. proposed a method of SRL based on Japan FrameNet [16]. They employed ME and SVM for argument identification and semantic role assignment. In order to train stochastic models, their method boosts annotated examples also.

For Chinese language, SRL methods that are successful on English are adopted to resolve Chinese SRL [17, 18]. [18] produced complete and systematic research on full parsing based methods. Their method divided SRL into three sub-tasks: 1) pruning with a heuristic rule, 2) Argument Identification (AI) to recognize arguments, and 3) Semantic Role Classification (SRC) to predict semantic types. For Tamil documents, [19] proposed SRL on Maximum Entropy Model and showed Evaluation phrase for SRL of Tamil texts. [20] presented an SRL system for Modern Standard Arabic. It showed the experiments on the Arabic Propbank data based on SVM and Kernel Mehods.

3 Myanmar Verb Frame Resource

Myanmar language is a free word order. It is very different from English language. Myanmar language is Subject Object Verb (SOV) order. [21] proposed Myanmar Verb Frame (MVF) as a lexical resource which represents the relationship between a predicate and its arguments of the Myanmar language. Myanmar verb frame files build together with example sentences annotated with semantic roles following PropBank guidelines. But, this system could not reproduce the same experience of PropBank. This system interested in designing Myanmar Verb Frame files in relatively independent modules to facilitate the collaborative construction of this resource.

Table 1 Semantic roles in MVF

Tag	Description
Arg0	Agent(usually the subject of a transitive verb)
Arg1	Patient(usually its direct object or the subject of a intransitive verb)
Arg2	Instrument, benefactive
Arg3	starting point
Arg4	ending point
Argm-loc	Locative
Argm-tmp	Temporal
Argm-mnr	Manner
Argm-cau	Cause
Argm-prp	Purpose
Argm-dir	Direction

Once PropBank guidelines and PropBank frames files are available for consultation, it is design to adopt a different approach: instead of firstly building frames files and Annotator´s Guidelines. Myanmar Verb Frame is started by annotating a corpus using English frames files and guidelines as model. Therefore, unlike PropBank, in this first phase it annotated only semantic role labels and not verb senses. In this way, there are identified language-specific aspects of SRL for Myanmar language. In this study, we worked with a set of 11 predicate-independent abstract semantic roles in Table 1.

4 SRL Method for Myanmar Sentences

Myanmar is a very different language from English in several respects relevant to the SRL task. Myanmar language exhibits rich morphology. Our SRL task is divided into two subtasks. Before performing SRL task, as preprocess step, Myanmar Earely Parser [22] use to syntax structure trees of input sentences.

4.1 Predicate-Arguments Identification

The first step of SRL is arguments to which semantic roles should be assigned are selected. In other words, constituents are identified in predicate-argument identification.

```
Begin
Input    : Parser tree of Input sentence.
Output: Target V, Array of Arg_Candidate
        Arg[]={}, New_arg={}, New_argList[]={}, i=0
Step 1: Find Target V in input tree.
Step 2: Find the sisters of Target V in the same level.
          Arg[]= sisters of Target V
Step 3: for each Arg[] do
        if(Arg[i]  ∈  preposition)
          New_arg+=Arg[i]
        Else
          New_argList[i]=New_arg+Arg[i]
        Return Target V.
        Return New_argList[].
End
```

4.2 Semantic Roles Mapping Algorithm for Arguments

The second task of our SRL method is to assign an appropriate semantic role to each constituent. From the first step, we got what is predicate (verb) in sentence and what are arguments. In this step, we find semantic roles of arguments

associated with predicates from MVF using following mapping algorithm for arguments with syntactic constituents.

```
Begin
Input: Predicate (V), Array of Argument (New_argList[]),
       Myanmar Verb Frames
Output: SRL Sentence
Step1:  Search Verb Frame for corresponding input predicate in Myanmar Verb Frame Resource.
Step 2: for each New_argList[] do
    {
    if(Argument of Verb Frame  ∈  Syntactic tag of
    New_argList[i])
      Arg_Candidate= Arg_Candidate+Argument of Verb Frame
    }
    SRL Sentence+= Argument Candidate
End
```

According the mapping algorithm, semantic role is indicated by a particular syntactic position (example. object of a particular preposition) as shown in Figure 1.

- Agent: subject (PREP_NOM)
- Patient: direct object (PREP_OBJ)
- Instrument: object of "with/by" (PREP_ACCURATION)
- Source: object of "from" (PREP_DEPATURE)
- Destination: object of "to" (PREP_ARRIVAL)

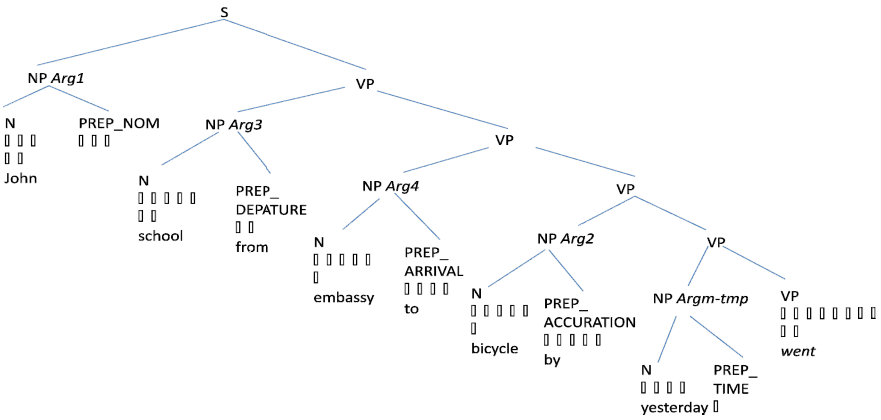


Fig. 1 Annotated Myanmar sentence corresponding to "John went to embassy from school by bicycle yesterday".

5 Experiment

Like English language, Myanmar language does not have Treebank. Therefore, we use parse trees of dataset sentences that passed through the Earsely Parser. The test set contains about 1,000 parsed sentences. Our SRL approach does not need training data and testing data. It does not use classifier. So, these parsed sentences are the input of our algorithms. The table 2 shows success rate of algorithms on the test set. This semantic role mapping algorithm performs on only main arguments in input sentences. The measurement for success rate on the test set can be as follows:

$$Success\ Rate = \frac{Number\ of\ Correctly\ Labeled\ Arguments}{Number\ of\ Total\ Arguments}$$

Table 2 Success Rate on the test set

Total Sentences	1,000
Total Main Arguments	2420
Total Modifier Arguments	390
Correctly Labeled Arguments	2080
Success Rate	74%

6 Conclusions and Future Work

This paper have presented first SRL method for Myanmar language that yields a global SRL over 70% by combining arguments identification and mapping algorithm unlike other language SRL systems. To assign semantic roles from MVF, mapping algorithm depends on the syntax structure of the sentences. The propositions in syntax structure of sentences are main points to label with semantic roles according to SRL method. This semantic role mapping algorithm with arguments still needs to improve for modifier arguments in sentences. This SRL approach need to improve and test a lot of test sets. Therefore, we will do the improvement of SRL approach as our future work. And then, we would like to use this SRL method for Myanmar text summarization system and other NLP applications.

References

1. Gildea, D., Jurafsky, D.: Automatic Labeling of Semantic Roles. Computational linguistics 245–28 (2002)
2. Gildea, D., Palmer, M.: The necessity of syntactic parsing for predicate argument recognition. In: Proceedings of ACL 2002, Philadelphia, USA (2002)
3. Gildea, D., Hockenmaier, J.: Identifying semantic roles using combinatory categorical grammar. In: Proceedings of ACL 2002, Philadelphia, USA (2002)

4. Fleischman, M., Kwon, N., Hovy, E.: Maximum Entropy Models for FrameNet Classification (2003)
5. Thompson, C.A., Levy, R., Manning, C.: A generative model for SRL. In: Proceedings of ECML 2003, Dubrovnik, Croatia (2003)
6. Surdeanu, M., Harabagiu, S., Williams, J., Aarseth, P.: Using predicate-argument structures for information extraction. In: Proceedings of ACL 2003, Sapporo, Japan (2003)
7. Hacioglu, K., Ward, W.: Target word detection and semantic role chunking using support vector machines (2003)
8. Pradhan, S., Hacioglu, K., Ward, W., Martin, J.H., Jurafsky, D.: Semantic role parsing: adding semantic structure to unstructured text. In: Proceedings of theInternational Conference on Data Mining (ICDM-2003), Melbourne, USA (2003)
9. Xue, N., Palmer, M.: Calibrating Features for SRL. EM-NLP (2004)
10. Liu, T., Che, W., Li, S., Hu, Y., Liu, H.: SRL System using Maximum Entropy Classifier. CoNLL (2005)
11. Gildea, D., Jurafsky, D.: Automatic labeling of semantic roles. Computational Linguistics **28**(3), 245–288 (2002)
12. Kwon, N., Fleischman, M., Hovy, E.: Framenet-based semantic parsing using maximum entropy models. In: Proceedings of COLING 2004, pp. 1233–1239 (2004)
13. Kingsbury, P., Palmer, M.: From tree-bank to propbank. In: Proceedings of LREC 2002 (2002)
14. Pradhan, S., Ward, W., Hacioglu, K., Martin, J.H., Jurafsky, D.: Shallow semantic parsing using support vectormachines. In: Proceedings of HLT/NAACL 2004 (2004)
15. Bejan, C.A., Moschitti, A., Morarescu, P., Nicolae, G., Harabagiu, S.: Semantic parsing based on FrameNet. In: SENSEVAL-3, Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text/ACL 2004, pp. 73–76, July 2004
16. Hizuka, S., Okamoto, H., Saito, H., Ohara, K.: SRL based on Japanese FrameNet (in Japanese). Journal of Natural Language Processing **14**(1), 43–66 (2007)
17. Sun, H., Jurafsky, D.: Shallow semantic parsing of chinese. In: Proceedings of NAACL 2004, Boston, USA (2004)
18. Xue, N.: Labeling Chinese predicates with semantic roles. Computational Linguistics **34**, 225–255 (2008)
19. Pandian, S.L., Geetha, T.V.: SRL for Tamil Documents. International Journal of Recent Trends in Engineering, **1**(1) (May 2009)
20. Diab, M., Moschitti, A., Pighin, D.: SRL Systems for Arabic using Kernel Methods. ACL (2008)
21. Naing, M.T., Thida, A.: Myanmar proposition bank: verb frame resource and an annotated corpus of semantic roles. In: Proceedings of ASEAN Community Knowledge Networks for the Economy, Society, Culture, and Environmental Stability, p. 42, May 2014
22. Phyu, S.L.: Development of Lexico-Conceptual Knowledge Resources and Syntax Analyzer for Myanmar Language. Ph.D. Thesis, University of Computer Studies, Mandalay, Myanmar (2013)

Text Document Clustering with Ontology Applying Modify Concept Weighting

Hmway Hmway Tar and Myint Myint Khaing

Abstract With the increasing amount of information, researchers in digital communities have witnessed the tremendous growth of publications. The overwhelming amount of information still makes it a time-consuming task. There are many of computer science and medical subject related documents cited on the Internet. Ontologies currently are hot topics in the area of Semantic Web. Ontologies can also help in addressing the problem of searching related entities, including research publications. The purpose of the system is to cluster the text documents based upon the ontology. The system is applying the modified concept weighting and become the extended version of the work that has been done before [8]. After the time passed the testing amount of data becomes lager and the challenges is the time complexity. To overcome this issue the system used the scoring method at the concept weighting stages to manage the time complexity. The experiments reveal that even the testing documents increased; the system may actually be able to produce useful result for text document clustering.

Keywords Ontology · Semantic Web · Text document clustering

1 Introduction

While the capabilities of today's Web directed towards the Semantic area, many research for the field of ontology become more interested area. With the booming of the Internet, the World Wide Web contains a billion of textual documents. This factor put the World Wide Web to urgent need for clustering method based on ontology which are developed for sharing, representing knowledge about specific domain. To explore and utilize the huge amount of text documents, many methods are developed to help users effectively navigate, summarize, and organize text

H.H. Tar · M.M. Khaing(✉)
University of Computer Studies, Pinlon, Myanmar
e-mail: {Hmwaytar34,myintkhaing06}@gmail.com

T.T. Zin et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 388,
DOI: 10.1007/978-3-319-23207-2_44

documents that is why clustering become an important factor. However, as more text documents are populated, many systems urgently need to rely on well model technologies such as Semantic Web.

Document Clustering become an essential technology with the popularity of the Internet. That also means that fast and high-quality document clustering technique play core topics. Text clustering is grouping semantically related items. This is also called unsupervised learning because no training data is available to help in deciding which group an item should belong to. Traditional clustering techniques depend only on term strength and document frequency which can be easily applied to clustering. This system also considers concept weight with the support of ontology.

Ontologies currently are hot topics in the area of Semantic Web. To effectively use that data and information this system applies ontology concepts to develop well defined model for data with well structure. This research is mainly concerned with the concept weighting and grouping algorithm by taking the advantages of the concepts of domain ontologies. Moreover, it is vital to have a reliable way to cluster massive amounts of text data. This method present a new way to mine documents and cluster them using ontology. One of the goals of the system is to clu ster text documents based on their concept weight similarity rather than keywords. This phase focuses on the introduction of the concept of semantic features weight similarity into the clustering methods. Text clustering algorithms have focused on the management of numerical and categorical data. However, in the last years, textual information has grown in importance. Proper processing of that kind of information within data mining methods requires at a semantic level. In the system's work, the concept of modify concept weighting is introduced to provide a formal framework for clustering documents. Available knowledge is formalized by means of ontology. Clustering cover approaches completely or partially relying on ontology. In the system values represent concepts weight rather than simple term weight. As a consequence, applying the results obtained in the first part of this research to the clustering processes should also have benefits on having a better identification of the clusters than non-semantic clustering. On the other hand, it has been proposed a method to include semantic features weight into an unsupervised clustering.

2 Background Theory

Text mining is a technique developed from data mining to analyze textual data especially unstructured (free text, abstract, etc). A text document is unclear, and according to [1, 2, 3]. Traditionally, ontology has been defined as the philosophical study of what exists: the study of kinds of entities in reality, and the relationships that these entities bear to one another. In the context of computer and information sciences, ontology defines a set of representational primitives with which to model a domain of knowledge or discourse. The representational primitives are typically classes (or sets), attributes (or properties), and relationships (or relations among class members). In computer and information science, ontology is a technical term denoting an artifact that is designed for a purpose, which is to enable the modeling of knowledge about some domain, real or imagined [4,5,6,7].

3 Problem Statement

Recently, researchers in digital communities have witnessed the tremendous growth of publications. Even though search engines on the Internet provide the efficient way for researchers to search publications of interests, the overwhelming amount of information still makes it a time-consuming task. Clustering, a technique used in any areas, is one way to facilitate this. Ontologies can also help in addressing the problem of searching related entities, including research publications.

Most of the existing text clustering methods use clustering techniques depends only on term strength and document frequency using TF-IDF formula in the document. But this method only considers the times which the words appear, while ignoring other factors which may impact the word weighs. And also this method is only a binary weighting method. This proposed system also considers concept weight for selecting the trait of the documents with the support of ontology so that the utility of ontology can be applied in clustering process. Moreover, this system wishes to utilize the ontology hierarchy structure it added the categorical information table before weighting phases. The previous system meets some obstacle when the applied document sets become lagers [8]. To overcome this issue this system incorporated with the grouping algorithm [9].

4 Proposed System

Searching the World Wild Web can be frustrating. Past studied have indicated that applying concept weight becomes biased towards some of the text documents when applying the tremendous testing data [8]. To counter act this problem we use grouping algorithm. Figure 1 provides main development of the system and also describes a detailed description of all the process that was taken out in this system.

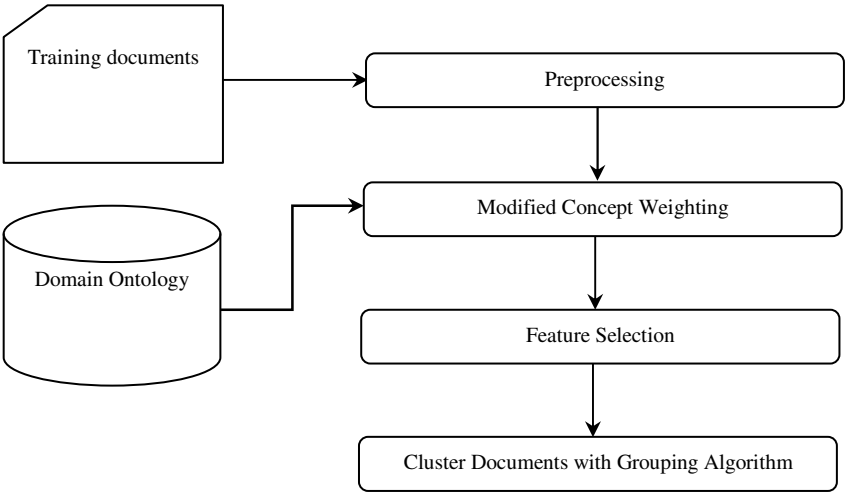


Fig. 1 Overview for the Proposed System

The implementation of the system consists of five parts. The first part is ontology creation; the second part is weighting calculation. The rest of the path is for clustering. The goal of this research is the development of a domain-specific ontology, which will be used for technology clustering. This section presents a detailed explanation of the system work, which can be used in combination with ontology concepts. The work has been based on information extracted and inferred from Google Search Engine relating with the dissertation papers about image processing domain, distributing system and medical domain. With the growing demands in the research and development community of image research, distributed system and medical field, it is necessary to capture concept hierarchical data in order to provide an efficient means and efficient model of these areas of research. Therefore, the system creates an ontology which can be queried to gain knowledge for this research area and discovery has been conceived. The basic steps in building ontology are straightforward. The system has explored the ontology construction using text documents as shown in Figure 1. This ontology is captured in the OWL DL (Ontology Web Language Description Logics) language and supported by the current ontology editors, valuators, and reasoners.

The text document collection is the initial stage for system. In the preprocessing stage, the document is transferred to a format suitable to the representation process. The textual information is stored in many kinds of machine readable form, such as PDF, DOC, PostScript, HTML, and XML and so on. However, there are still a lot of papers stored in the plain pdf format. After the text document are collected from Google search engine, the abstract of the paper is elective from those pdf file and transformed into TXT format and maintained in the text files. After that phase, the system removes the stop words and stemming on the extracted text document. The stop-words are high frequent words that carry no information (i.e. pronouns, prepositions, conjunctions etc).

In weighting phase the system calculate the weight of the concept as [8] as shown in below

$$W = \text{Length} \times \text{Frequency} \times \text{Correlation Coefficient} + \text{Probability of concept} \quad (1)$$

where W is the weight of keywords, Length is the depth of concept in the otology. Frequency is the times which count the words appear in the document, and if the concept is in the ontology Correlation Coefficient is taken as 1 and otherwise 0. Probability is based on the probability of the concept in the document. Finally the total Score is calculated by equation (2). Generally, every text should have a higher semantic similarity score with the texts from its group than the texts from different groups [10]. There are a few rare cases where this assumption could fail. One case is that the semantic similarity score does not reflect the relationships among the texts. Moreover, it is not always obvious what should be the attributes of the items and what should be their values. For example, for text documents, often words contained in them are used as the attributes, but, consecutive sequence alphabets in the document or consecutive sequence of two words, called bigrams, have also been used. The system assigned values to main attributes that will

specify obviously the representation of the items. The system used fixed score for some attribute contained in the texts by calculating score2. This can be used to weight the most useful of the attributes for the particular task. The score2 are pre-assign value with highest score for some words. The score2 are the fixed value according to the words that are found in the text documents as follows:

If the word found in the documents is *image*
Then the score2 will be assigned 10
If the word found in the documents is *distributed*
Then the score2 will be assigned 100
If the word found in the documents is *medical*
Then the score2 will be assigned as 300
If the word found in the documents is *intelligence*
Then the score2 will be assigned as 1000

$$Score = W + score2 \tag{2}$$

After calculating the *Score*, the last state is actionable. The basic idea is that each text could gather its most related texts to form an initial group, Yllias Chali decide which groups have more strength over other groups, make the stronger groups as final clusters, and use them to bring any possible texts to their clusters. First, Yllias Chali's system uses each text as a leading text (Tl) to form a cluster. To do this, they put all the texts which have a score greater than the high-threshold with Tl into one group and add each score to the group's total score. By doing this for all texts, they will have N possible different groups with different entries and group scores, where N is the number of the total texts in the set. Next, they select the final clusters from those N groups. They arrange a ll the groups by their scores in a non-increasing order. They choose the group with the highest score and check if any text in this group has been clustered to the existing final clusters or not. If not more than 2 texts are overlapping with the final clusters, then their algorithm take this group as a final cluster, and remove the overlapping texts from other final clusters. Yllias Chali's stated that the process the group with the next highest score in the same way until the groups' entries are less than 4. For those groups, they would first try to insert their texts into the existing final clusters if they can fit in one of them. Otherwise, they will let them go to the leftover cluster which holds all the texts that do not belong to any final clusters. After the concept weighting phase we apply the grouping algorithm. The following is the pseudocode for the grouping algorithm , Yllias Chali applied in their system:

// Get the Initial Clusters
For each text t i
Construct a text cluster including all the texts (t j) which score (ti,tj)>=high threshold;
Compute the total score of the text cluster;
Find out its neighbor with maximum relation score;
End For

```
//Build the final clusters
  Sort the clusters by their total score in non-increasing order;
  For each cluster g i in the sorted clusters
    If member gi >3 and overlap-mean gi <=2
      Take g i as a final cluster c i;
      Mark all the texts in ci as clustered;
    Else
      Skip to process next cluster;
    End If
  End For

//Process the leftover texts and insert them into one of the final clusters
For each text t j
  If t j has not been clustered
    Find cluster c i with the highest score (ci, tj);
  If the average-score (ci, tj) >= low- threshold
    Put tj into cluster ci;
 Else If the max score neighbor tm of tj is in ck  Put tj into cluster ck;
 Else
  Put tj into the final leftover cluster;
End if
End if
End For
Output the final clusters and the final leftover cluster;
```

5 Experimental Results

The proposed system has been tested with four test cases. The experiments in this section are conducted on the papers that are downloaded from the Google. The system downloaded 2000 papers from the Google Search track of recent World Wide Web conference websites. These 2000 test documents came from three sub-categories (types) of *Image* documents, *Distributed System* documents and *Medical* and *Aritificial Intelligence* related documents respectively. Table 1 shows the results of the four dataset's statistical relatedness analysis measures using precision and recall rate. As expected, the highest scoring produce the highest precision, which shows that the system score is a good measure of the degree of relevancy between concepts and the documents. The performance of the method is influenced by a number of factors. The CPU requirements for the experiments described above are of the order of 2-3 hours. The memory requirements are quite excessive, and there is a trade-off between the number of abstracts (instances) and the number of concepts (features). Moreover the performance time is rapidly increased as when using ontology time is the one of the issue for many applications.

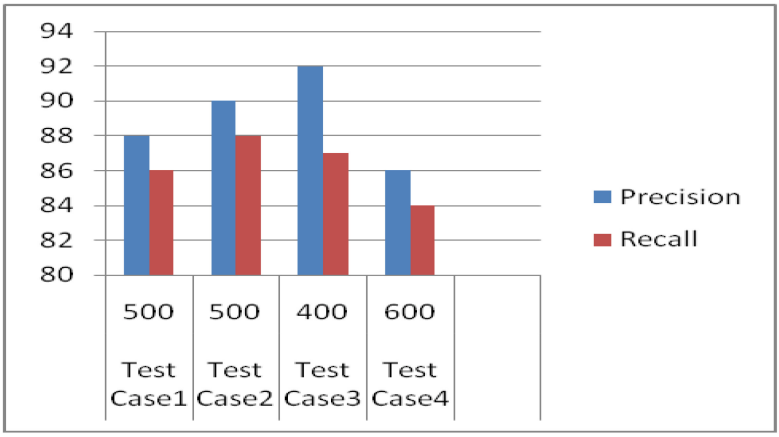


Fig. 2 Testing Results

6 Performance Analysis

The performance of the method is influenced by a number of factors. The system processing time may be slow degradation caused by the continuous growth of the ontology size and sudden improvement gains due to more successful arrangements of concepts and clustering methods has be seen. The CPU requirements for the experiments described above are of the order of 2-3 hours. The memory requirements are quite excessive, and there is need to be trade-off between the number of documents and the number of concepts if the processing time needs to be decreased. The concept/documents ratio may have to be reduced for very large-scale experiments that influence the ontology. The system performance also degrades if the number of documents is increased and if the ontology size is growing.

From the results of this research it has been shown that the idea of using ontology to represent the clustering document in place of its concept weighting, as conventionally is used, is a sound principle under certain conditions. The dominating condition with regard to the test dataset and the testing methods were the length of the article i.e. the number of words occurring in the document. The achieved results indicate that as the average number of words in a document corpus increases the less of an impact that the ontology methods. In conclusion, the ontology-based approach performed as good even it have some minor differences than traditional clustering method and statistically method.

7 Conclusion

The main aim of the work is the development of a methodology able to exploit ontological computing when used in clustering methods, called ontology-based clustering. Thw work is a contribution in the field of ontology, in which the

system has studied how domain knowledge can been exploited before the clustering process. Moreover, these approaches do not attempt to interpret the conceptual meaning of textual terms, which is crucial in many applications related to textual data. The system has focused on applying semantic issue. Moreover, the system applied medical domain related papers to test whether the system can give accurate cluster and also required medical knowledge is acquired from the medical knowledge obtained from our second author [8]. As expected as earlier, the experimental results illustrating the effectiveness of the technique. Therefore the system should extend this technique more statistically for further experiments to conduct more accurate text document clustering because the author interests area is the clustering system which can catch up with Google clustering methods which was hit 90% in this year for all domain.

The system observed that the clustering results are affected by the degree of completeness of the ontology. For the case of evaluating, the four test cases were analyzed. The system also observed that the more accurate the ontology the more accurate the clustering results. Successful results were obtained with the domain ontology. In that sense, the proposed system clustering approach can be used in several tasks and domains such as electronic commerce (e.g. grouping similar products or to obtain a characterization of users), medicine (e.g. clustering of electronic health records), tourism (recommending tourist destinations to users), even in privacy preserving.

References

1. Jursic, M., Lavrac, N.: Fuzzy Clustering Of Documents. Department of Knowledge Discovery
2. SteinBach, M., Karypis, G., Kumar, V.: A comparison of document clustering techniques. In: KDD Workshop on Text Mining (2000)
3. Reinberger, M.-L., Spyns, P.: Discovering knowledge in texts for the learning of DOGMA-inspired ontologies. In: Proceedings of ECAI 2004 Workshop on Ontology Learning and Population (2004)
4. Noy, N.F., McGuiness, D.L.: Ontology Development 101: A Guide to Creating Your First Ontology. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880 (March 2001)
5. Bechhofer, S., Horrocks, I., Goble, C.A., Stevens, R.: OilEd: a reason-able ontology editor for the semantic web. In: Baader, F., Brewka, G., Eiter, T. (eds.) KI 2001. LNCS (LNAI), vol. 2174, pp. 396–408. Springer, Heidelberg (2001)
6. Karapiperis, S., Apostolou, D.: Consensus Building in Collaborative Ontology Engineering Processes. Journal of Universal Knowledge Management **1**(3), 199–216 (2006)
7. Berners-Lee, T.: Weaving the Web. Harper, San Francisco, HarperCollins Publishers, New York (1999)
8. Tar, H.H., Oo, P.P.: Ontology employment in Text Document Clustering combined with Grouping Algorithm. IJAIS **6** (2013)
9. Chali, Y., Noureddine, S.: Document Clustering with Grouping and Chaining Algorithms. University of Lethbridge
10. Pantel, P., Lin, D.: Document clustering with committees. In: Proceedings of the ACM SIGIR 2002, Finland (2002)

Ontology Based Comparative Sentence and Relation Mining for Sentiment Classification

Myat Su Wai, May Aye Chan Aung and Than Nwe Aung

Abstract Due to the rapid expansion of the internet, business through e-commerce has become popular. Many products are being sold on the internet and the merchants selling the products ask their customers to write reviews about the products that they have purchased. Opinion mining and sentiment classification are not only technically challenging because of the need for natural language processing, but also very useful in practice. In this study, ontology based compararive sentence and relation mining for sentiment classification in mobile phone (product) reviews are studied. POS taggers are used to tag sentiment words in the input sentences. In this study, Naive Bayes classifier is also used for sentiment classification. Moreover, the comparison between with ontology and without ontology are aiso described. This study is very useful for manufacturers and customers in E-commerce Sites, Review Sites, Blog etc.

Keywords Sentiment classification · Ontology · Naïve Bayes classifier

1 Introduction

Due to the rapid expansion of the internet, business through e-commerce has become popular. Many products are being sold on the internet and the merchants selling the products ask their customers to write reviews about the products that they have purchased. This is the reason behind the abnormal increment of the number of reviews on websites.

Opinion mining and sentiment classification are not only technically challenging because of the need for natural language processing, but also very useful in

M.S. Wai(✉) · M.A.C. Aung · T.N. Aung
University of Computer Studies, Mandalay, Myanmar
e-mail: {missmyatsuwai,mayayechanaung,mdytna}@gmail.com

T.T. Zin et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 388,
DOI: 10.1007/978-3-319-23207-2_45

practice. For example, businesses always want to find public or consumer opinions on their products and services. Potential customers also want to know the opinions of existing users before they use a service or purchase a product. Moreover, opinion mining can also provide valuable information for placing advertisements in Web pages. In a page, if people express positive opinions or sentiments on a product, it may be a good idea to place an ad of the product. And if people express negative opinions about the product, it is probably not wise to place an ad of the product.[13] [7]. The Web has dramatically changed the way that people express their opinions. They can now post reviews of prod-ucts at merchant sites and express their views on almost anything in Internet forums, discussion groups, blogs, etc., which are commonly called the user generated content or user generated media.

2 Related Works

Several techniques are used for the opinion mining tasks. To extract opinions, machine learning method and lexical pattern extraction methods are used by many researchers [3]. The researches about comparative sentences are mainly in two fields, linguistics and computational linguistics. Researchers in linguistics are concerned with semantics and syntax of comparative sentences, instead of the automatic recognition technology. Songbo Tan, Zhang [8] summarized the various classification systems of comparative sentences in modern Chinese documents. It can be seen that there are not uniform opinions how to classify comparative sentences in Chinese study. Khin Phyu Phyu Shein et al., [6] described on the Internet there are lots of contents that opinions or sentiments about an object such as reviews about music, movie, software, products and books etc. The aim of sentiment classification is to extract the feature on which reviewer express their emotion or feeling and identify them as positive, negative or neutral. In this paper, proposed model was the combination of Support Vector Machine with Natural Language Processing techniques. Ontology based on Formal Concept Analysis design for classifying the software reviews is negative, positive or neutral. It is proposed model that mainly focus is on feature level sentiment classification. Three main parts in this approach are: assigning the POS tags, identifying domain related features and classifying the sentiment words [5].

In computational linguistics, Huang et al [4], applied several supervised machine learning methods to classify a Chinese sentence into either “comparative” or not. Song et al. constructed a Chinese comparative pattern data-base and used it to recognize comparative sentences. Alaa combined POS (Part of Speech) tags and several learning methods to extract comparative statements in Arabic [8].

A comparative relation defined in includes compared entities, compared features, a comparative keyword and a comparative type. Jindal and Liu extracted comparative relations using a new type of rules called LSR (Label Sequential Rules). Xu et al built a graphical model based on two-level CRF to recognize comparative relations and the directions of relations on mobile phone review data. Taras Zagibalov and John Carroll [9] constructed Automatic Seed Word Selection

for Unsupervised Sentiment Classification and used hybrid comparative patterns to label compared entities and compared features for Chinese comparative sentences. WeiWei and Gulla [10], proposed a HL-SOT approach to labeling product attributes and their associated sentiments in product reviews by a Hierarchical Learning (HL) process with a defined Sentiment Ontology Tree (SOT). However, the SOT is manually constructed. It is an effort of human beings that what attributes need to be included in SOT of the product and how to structure these attributes in the SOT. The sizes and structures of the SOT constructed by different individuals may vary [14]. This paper refers to ontology based comparative patterns to recognize comparative sentences and relations for sentiment classification with accurate results.

3 POS Tagging

Part-of-speech (POS) tagging is useful to our subsequent discussion and also the proposed techniques. In grammar, part-of-speech of a word is a linguistic category defined by its syntactic or morphological behavior. Common POS categories are: noun, verb, adjective, adverb, pronoun, preposition, conjunction and interjection. Then, there are many categories which arise from different forms of these categories. In this work, we use Brill's Tagger (Brill 1992). Important POS tags to this work and their categories are:

NN: Noun, NNP: Proper Noun, PRP: Pronoun, VBZ: Verb, present tense, 3rd person singular, JJR: Comparative Adjective, JJS: Superlative Adjective, RBR: Comparative Adverb, RBS: Superlative Adverb.

Although JJR, JJS, RBR, and RBS tags represent comparatives, many sentences containing such tags are not comparisons. Many sentences that do not contain any of these tags may be comparisons. Thus, we cannot solely use these tags to identify comparative sentences [11].

4 Sentiment Classification

Given a set of evaluative texts D, a sentiment classifier classifies each document into one of the two classes, positive and negative. Positive means that document d expresses a positive opinion. Negative means that document d expresses a negative opinion [1] [2].

The main application of sentiment classification is to give a quick determination of the prevailing opinion on an object. The task is similar but also different from classic topic-based text classification, which classifies documents into predefined topic classes, e.g., politics, science, sports, etc.

In topic-based classification, topic related words are important. However, in sentiment classification, topic-related words are unimportant. Instead, sentiment words that indicate positive or negative opinions are important, e.g., great, excellent, amazing, horrible, bad, worst, etc [1].

4.1 Comparative Sentence and Relation Mining

Directly expressing positive or negative opinions on an object is only one form of evaluation. Comparing the object with some other similar objects is another. Comparison is a more convincing way of evaluation. Comparisons are related to but also different from typical opinions. They have different semantic meanings and different syntactic forms. Comparisons may be subjective or objective [1].

A comparative sentence is a sentence that expresses a relation based on similarities or differences of more than one object. The comparison in a comparative sentence is usually expressed using the comparative or the superlative form of an adjective or adverb [1]. The comparative sentence is used to state that one thing has more (bigger, smaller) "value" than the other. The superlative is used to say that one thing has the most (the biggest, the smallest) "value". The structure of a comparative sentence consists normally of the stem of an adjective or adverb, plus the suffix -er, or the modifier "more" or "less" before the adjective or adverb. For example, in "Sony phone is better than Nokia", "better" is the comparative form of the adjective "good". The structure of a superlative sentence consists of the stem of an adjective or adverb, plus the suffix -est, or the modifier "most" or "least" before the adjective or adverb. In "Sony is the best in the phones", "best" is the superlative form of the adjective "good". A comparison can be between two or more objects, groups of objects, one object and the rest of the objects.

4.2 Types of Important Comparisons

Types of important comparisons can be classified into four main types. The first three types are gradable comparisons and the last one is the non-gradable comparison. The gradable types are defined based on the relationships of greater or less than, equal to, and greater or less than all others.

1. Non-equal gradable comparisons: Relations of the type greater or less than that express an ordering of some objects with regard to some of their features, e.g., "the Samsung is faster than that of Huawei". This type also includes user preferences, e.g., "I prefer Samsung to Huawei".
2. Equative comparisons: Relations of the type equal to that state two objects are equal with respect to some of their features, e.g., "the picture quality of Sony is as good as that of Samsung"
3. Superlative comparisons: Relations of the type greater or less than all others that rank one object over all others, e.g., "the Sony is the fastest".
4. Non-gradable comparisons: Sentences that compare features of two or more objects, but do not grade them.

The first three types of comparative are called gradable comparatives [1]. This work only focuses on these three types. For simplicity, from now on we use comparative sentences and gradable comparative sentences interchangeably. Note that in a long comparative sentence, there can be multiple relations separated by delimiters such as commas "," and conjunctions such as "and" and "but".

5 Ontology Based Approach to Sentiment Classification

Ontology has been defined as the specialization of the conceptualization. The main aim of ontology is to provide knowledge about specific domains that are understandable by both the computers and developers [9].

Ontology can be used to describe the domain and to reason about the entities with-in that domain. To construct ontology the knowledge about particular field of interest is needed. Ontology designed for one domain cannot be applied to another one. [8] [9]. Ontology-based approach can be used in classification of opinions and in feature-based opinion mining. In both cases ontology can be used in a different ways. In most popular approach single opinion can be presented as an instance of ontology. The comparison analysis of those instances should be conducted in classification of collected opinions. The polarities of opinions with the same subject can be aggregated to the overall sentiment to the product or service [6].

The ontology as graph-like construction makes feature based opinion mining easier to conduct. The main characteristics of the subject of opinion can be presented as its attributes in the ontology. Then the polarity of each feature must be determined either for single opinion or for the whole set of opinions. In a special case of this approach for every node two additional leafs representing positive and negative sentiment are added [12].

5.1 Creating Mobile Phone Ontology

In phone ontology, “Phone” is a main class and it has subclasses and relations with them .We use HasA relation and IsA relation between these class and subclasses. And we also use other relations such as include and provide. We construct Mobile phone ontology by using Protégé 4.3 as shown in figure.

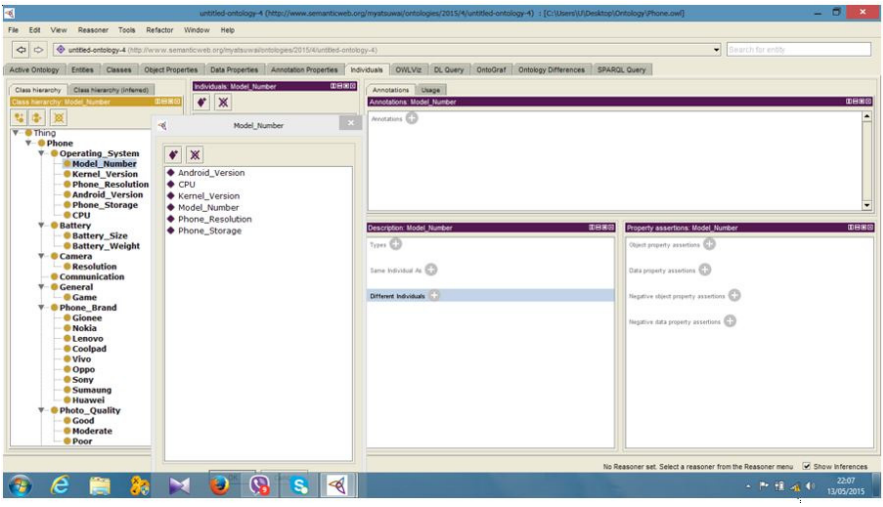


Fig. 1 Phone Ontology Creation for Phone Domain that are used in Sentiment Classification

6 Naïve Bayes Classifier

Most of the algorithms for sentiment analysis are based on a classifier trained using a collection of annotated text data. Before training, data is preprocessed so as to extract the main features. Some classification methods have been proposed: Naive Bayes, Support Vector Machines, K- Nearest Neighbors, etc. However, according to Go et al., 2009, it is not clear that which of these classification strategies is the more appropriate to perform sentiment analysis.

We decided to use a classification strategy based on Naive Bayes (NB) because it is a simple and intuitive method whose performance is similar to other approaches. NB combines efficiency (optimal time performance) with reasonable accuracy. If the main features are the tokens extracted from texts, it is evident that they cannot be considered as independent, since words co-occurring in a text are somehow linked by different types of syntactic and semantic dependencies. However, even if NB produces an oversimplified model, its classification decisions are surprisingly accurate [10].

7 Discussions

There is only one relation in a sentence. In practice, this is violated only in a very small number of cases. Entities or features are nouns (includes nouns, plural nouns and proper nouns) and pronouns. These cover most cases. However, a feature can sometimes be a noun used in its verb form or some action described as a verb (e.g., “Sony costs more”; “costs” is a verb and a feature). Such comparisons are adverbial comparisons and are not considered. To evaluate sentiment classification system, we use the customer review of mobile phone form multi-domain dataset which were grouped into positive and negative categories for content analysis and tested and compared with the manually tagged set of review datasets. The results and comparisons are shown in the following table.

Table 1 Comparison Between with the use of ontology and without ontology

	Without the use of domain ontology	Proposed Approach (With ontology)
Positive Sentiment	Accuracy: 72% Recall : 90%	Accuracy: 80% Recall : 90%
Negative Sentiment	Accuracy: 64% Recall : 85%	Accuracy: 76% Recall : 85%
Ambiguous Sentiment	Accuracy: 65% Recall : 90%	Accuracy: 80% Recall : 90%
Neutral Sentiment	Accuracy: 60% Recall : 90%	Accuracy: 75% Recall : 90%

8 Conclusion

Opinion mining and sentiment classification are not only technically challenging because of the need for natural language processing, but also very useful in practice. Opinion mining can be served in the field of Information search & Retrieval. In opinion mining Determining sentiments seems to be easier, determining objects and their corresponding features is harder. Combining both the task is very tedious and also accuracy is the problem. In this study, ontology based comparative sentence and relation min-ing for sentiment classification in mobile phone (product) review have been proposed. Naïve Bayes classifier is used to classify word of comparative sentences and relations from ontology. According to study, ontology based approach is better in accuracy and recall than other approaches, with ontology.

References

1. Liu, B.: Web Data Mining. Springer, Department of Computer Science, University of Illinois, Chicago
2. Liu, B.: Sentiment analysis and subjectivity. In: Indurkhya, N., Damerau, F.J. (eds.) Hand book of Natural Language Processing, 2nd edn (2010)
3. Pang, B., Lee, L.: Using very simple statistics for review search: an exploration. In: Proceedings of the International Conference on Computational Linguistics (COLING). Poster paper (2008)
4. Hanshi, W., Xinhui, N., Lizhen, L.: A Fuzzy Domain Sentiment Ontology based Opinion Mining Approach for Chinese Online Product Reviews. Journal of Computers **8**(9) (September 2013)
5. Shein, K.P.P.: Ontology based combined approach for, sentiment classification. In: Proceedings Of The 3rd International Conference On Communications and Information Technology. ISSN: 1790-5109
6. Shein, K.P.P., Nyunt, T.T.S.: Sentiment classification based on ontology and SVM classifier. In: 2010 Second International Conference on Communication Software and Networks. IEEE (2010). doi:10.1109/ICCSN.2010.35
7. Dave, K., Lawrence, S., Pennock, D.M.: Mining the peanut gallery: opinion ex-traction and semantic classification of product reviews. In: Proceedings of WWW, pp. 519–528 (2003)
8. Mukras, R, Carroll, J.: A comparison of machine learning techniques applied to sentiment classification (2004)
9. Wai, M.S.: Study on Ontology Markup Languages. International Conference on Computer Application, UCSY (2015)
10. Pang, B., et al.: Sentiment classification using machine learning methods. In: EMNLP-2002
11. Padmaja, S., Sameen Fatima, S.: Opinion Mining and Sentiment Analysis – An Assessment of Peoples' Belief: A Survey. International Journal of Ad hoc, Sensor & Ubiquitous Computing (IJASUC) **4**(1) (February 2013)
12. Tan, S., Zhang, J.: An empirical study of sentiment analysis for Chinese documents. Expert Systems with applications **34**(4), 2622–2629 (2008)

13. Zagibalov, T., Carroll, J.: Automatic Seed Word Selection for Unsupervised Sentiment Classification of Chinese Text. Expert Systems with Applications **34**(4), 2622–2629
14. Wei, W., Gulla, J.A.: Sentiment learning on product reviews via sentiment ontology tree. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 404–413. Uppsala, Sweden, July 11-16, 2010

Word Boundary Identification for Myanmar Text Using Conditional Random Fields

Win Pa Pa, Ye Kyaw Thu, Andrew Finch and Eiichiro Sumita

Abstract This paper examines the effectiveness of conditional random fields (CRFs) when used to identify Myanmar word boundaries within a supervised framework. Existing approaches are based on the method of maximum matching which appears to suffer from problems relating to the manner in which Myanmar words are composed. In our experiments, the CRF approach is compared against a baseline based on maximum matching using dictionaries from the Myanmar Language Commission Dictionary (word only) and a manually segmented subset of the BTEC1 corpus. The experimental results show that the CRF model is able to achieve considerably higher F-scores on the segmentation task than the baseline, even when the baseline is allowed to use words from the test data in its dictionary.

1 Introduction

In the writing systems of many Asian languages, such as Myanmar, Chinese, Japanese and Thai, words are not delimited by spaces. There are no blanks in Myanmar text for word boundaries. Determining the word boundaries, and thus tokenizing the text, is usually one of the first necessary processing steps for Natural Language Processing (NLP) applications. Segmenting Myanmar text is not a trivial task since Myanmar text is composed of words consisting of one or more syllables, and one or more characters can also represent a syllable. Therefore word segmentation is an issue for natural language processing. It may also be necessary to allow multiple correct

W.P. Pa(✉)
Natural Language Processing Lab, University of Computer Studies, Yangon, Myanmar
e-mail: winpapa@ucsy.edu.mm

Y.K. Thu · A. Finch · E. Sumita
Multilingual Translation Lab, National Institute of Information
and Communications Technology, Kyoto, Japan
e-mail: {yekyawthu,andrew.finch,eiichiro.sumita}@nict.go.jp

T.T. Zin et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 388,
DOI: 10.1007/978-3-319-23207-2_46

segmentations of the same text, depending on the requirements of further processing steps. Word segmentation is a necessary prerequisite for higher level language analysis including named entity recognition and syntactic parsing that are used in many NLP applications such as machine translation, automatic speech recognition and information retrieval. Word segmentation is considered to be an important first step for natural language processing tasks.

2 Related Work

The problems of Myanmar word segmentation have been analyzed and different approaches have been developed to achieve different goals.

[1] proposed a hybrid approach that works by longest matching on syllable-segmented sentences. Their probabilistic model used a lexicon of 20,000 words from a Myanmar grammar [9] and achieved 0.755 precision. In their method of longest matching the known words from the dictionary are first segmented and subsequently an n-gram model predicts the segmentation of the unknown words. The principal problem of this approach stems from the ambiguity in the longest matching process, since words can be formed in more than one way.

[5] proposed a word segmentation approach that involved rule-based syllable segmentation and dictionary-based statistical syllable merging using a dictionary of about 30,000 words provided by the Myanmar NLP team of Myanmar Computer Federation. Their approach achieved 100% syllable accuracy and 98.94% precision, 99.05% recall and 98.99% F-score on their word segmentation task.

[6] proposed a 2-step longest matching approach. The first step, was syllable segmentation, in the second step left-to-right syllable longest matching forward segmentation was performed. A 2 million sentence monolingual Myanmar corpus and an 80K sentence English-Myanmar parallel corpus, and lists of stop words, syllables and words were used in the decision process for annotating word boundaries. This approach employed a similar longest matching strategy to [1], and as a consequence also suffers the same problem of ambiguity mentioned earlier.

[3] studied word segmentation in the context of statistical machine translation using 7 different schemes: manually annotated segmentation; character breaking; syllable breaking; syllable breaking + maximum matching; unsupervised word segmentation; syllable breaking + maximum matching + unsupervised word segmentation; and supervised word segmentation. Their study examined the effect of segmentation on the following language pairs: Myanmar to Japanese, Korean, Hindi, Thai, Chinese and Arabic languages. They proposed a new algorithm for Myanmar syllable breaking that achieved 100% accuracy and that can be easily adapted to related Asian syllabic languages such as Khmer, Laos, and Nepali. Their proposed unsupervised segmentation approach did not exceed the performance of the simpler maximum matching approach, and one plausible cause is the lack of data. In this work we focused on a supervised approach which we expected to perform well training on a small amount of human segmented data.

3 Segmentation

This section describes the segmentation methods that were used for the experiments in both the pre-processing stage and the word segmentation stage. The word segmentation was done from both character segmented data and syllable segmented data using CRFs.

3.1 Character Segmentation

The character segmentation pre-processing step trivially segmented the Myanmar sentence into a sequence of graphemes represented by the Unicode characters.

3.2 Syllable Breaking

Syllable breaking is a necessary step for Myanmar word segmentation, since most Myanmar words are composed of multiple syllables and most of the syllables are composed of more than one character. We used the algorithm of [3] for syllable breaking. There are three general rules to break Myanmar syllables from Unicode input text where a consonant is followed by dependent vowels and other symbols. For example, the word ကျောင်း (school) can be decomposed as: က+ျ+ော+ာ+င်+း. Here, the medial consonant ျ (Ya), vowel sign ော် (E), vowel sign ာ (Aa) follow consonant က (Ka) and sign ် (Asat) and sign း (Visarga) follow syllable final consonant (Nga). The exception to this combination rule is Kinzi, the conjunct form of U+1004 + Myanmar letter Nga, (e.g. င+်+ ္+ ဂ for င် in အင်္ဂလိပ် (English)) that precedes the consonant.

The first rule puts a word break in front of consonants, independent vowels, numbers and symbol characters. The second rule removes any word breaks that are in front of subscript consonants, Kinzi characters, and consonant + Asat characters. Break points for special cases such as syllable combinations of loan words (e.g. ဂျော့ချ်), that is the transliteration of "George", Pali words, phonologic segmentation (e.g. တက် က သိုလ်) and orthographic segmentation (e.g. တက္ကသိုလ်). In experiments for these rules with a 27,747 word dictionary the approach was able to achieve 100% precision and recall.

Fig. 1 An example of syllable breaking for a sentence.

3.3 Maximum Matching

Maximum matching is one of the most popular structural segmentation algorithms and it is often used as a baseline method in word segmentation [7]. This method segments using segments chosen from a dictionary. The method strives to segment using the longest possible segments. It is a greedy algorithm and is therefore sub-optimal. The segmentation process may start from either end of the sequences.

3.4 Conditional Random Fields

Linear-chain conditional random fields (CRFs) [4] are models that consider dependencies among the predicted segmentation labels that are inherent in the state transitions of finite state sequence models and can incorporate domain knowledge effectively into segmentation. Unlike heuristic methods, they are principled probabilistic finite state models on which exact inference over sequences can be efficiently performed. The model computes the following probability of a label sequence $\mathbf{Y} = \{y_1, \ldots, y_T\}$ of a particular character string $\mathbf{W} = \{w_1, \ldots, w_T\}$.

$$P_{\boldsymbol{\lambda}}(\mathbf{Y}|\mathbf{W}) = \frac{1}{Z(\mathbf{W})} exp(\sum_{t=1}^{T} \sum_{k=1}^{|\boldsymbol{\lambda}|} \lambda_k f_k(y_{t-1}, \mathbf{W}, t)) \quad (1)$$

where $Z(\mathbf{W})$ is a normalization term, f_k is a feature function, and $\boldsymbol{\lambda}$ is a feature weight vector.

We used the CRF++ toolkit[10] to build the CRF models. The feature set used in the models (up to character/syllable tri-grams) was as follows (where t is the index of the character/syllable being labeled):

1. Character/syllable unigrams: $\{w_{t-2}, w_{t-1}, w_t, w_{t+1}, w_{t+2}\}$
2. Character/syllable bigrams: $\{(w_{t-1}, w_t), (w_t, w_{t+1})\}$
3. Character/syllable trigrams: $\{(w_{t-2}, w_{t-1}, w_t), (w_{t-1}, w_t, w_{t+1}), (w_t, w_{t+1}, w_{t+2})\}$

These n-grams were combined with label unigrams and bigrams to produce the feature set for the model.

4 Experiments

4.1 Data Setup

The CRF models were trained using a training set selected from a manually segmented 50,000-sentence subset of the Basic Travel Expression (BTEC1) corpus [2]. We ran four maximum matching experiments, drawing from two different dictionaries. The first dictionary was the 26,413-word Myanmar Language Commission (MLC)[8] dictionary; the second dictionary contained the first, and also included all

9,475 of the segments from the manually annotated corpus used to train the CRF models.

The experiments were performed using 10-fold jackknifing of the manually segmented BTEC1 50,000 sentences, therefore a test set of 5,000 sentences was used for each fold. A closed test for maximum matching was conducted with the larger dictionary the in order to obtain an approximate upper bound for the method using the available data. There experimental results report the average statistics over all 10 folds together with their standard errors.

4.2 Training with CRFs

The CRF models were trained on two different segmentations of Myanmar, character and syllable. For each character and syllable model, four separate models that used different tag sets were trained. These four tag sets were: {4,5}, {1,4,5}, {1,2,4,5} and {1,2,3,4,5} using the tag number notation in Table 1.

Examples of segmentation annotated using the four different tag sets are given in Fig. 2.

The meaning of the example sentence in Fig. 2 is: "**The weather is very fine**". It contains 8 syllables, tagged with all four tag sets in decreasing order of tag set size from top to bottom. It can be segmented into 4 segments, actually three words, a noun, an adverb and a verb. The first four syllables becomes a noun that means "**the weather**", the fourth and fifth syllables form an adverb meaning "**very**" and the

Table 1 List of segmentation tags.

Tag number	Tag	Position
1	<	The first syllable/character in a word
2	>	The second last syllable/character in a word
3	+	Represents both < and >
4	-	Others
5	|	Final syllable/character in a word

Table 2 The four tag sets used for segmentation.

Number of tags	Tag set
2	- |
3	< - |
4	< > - |
5	< > + - |

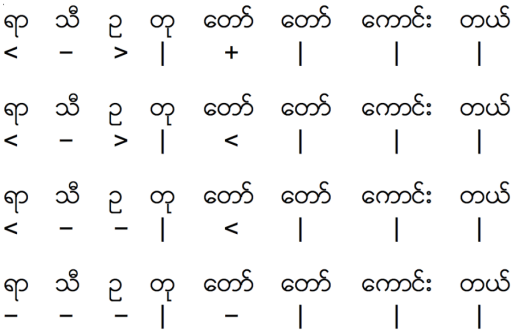


Fig. 2 Syllable tagging with different tags set.

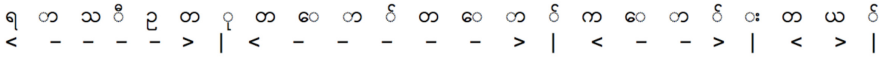


Fig. 3 Character tagging with 4 tags.

remaining two syllables form a verb meaning "**fine**". The verb is composed of two segments that are the root word and its suffix.

Fig. 3 gives an example of how the same sentence can be tagged at the character level with the 4 tags {1, 2, 4, 5}.

4.3 Evaluation Criteria

The segmentation performance of maximum matching and CRF models was measured using the commonly used precision (Equation 3), recall (Equation 4), and F-score (Equation 2) defined as follows.

$$\text{F-score} = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{2}$$

$$Precision = \frac{\#of\ correct\ tokens}{\#of\ tokens\ in\ test\ corpus} \tag{3}$$

$$Recall = \frac{\#of\ correct\ tokens}{\#of\ tokens\ in\ system\ output} \tag{4}$$

4.4 Results and Discussion

Table.3 gives the results on using various sizes of tag set in the CRF model. It is clear from the results that there were almost no differences in the performance of each of the systems. Therefore, for the remainder of the experiments we arbitrarily chose to use the largest tag set.

Table 3 Word segmentation performance (with standard errors) using different tag sets with CRF models.

Tagging Method	Character			Syllable		
	Precision	Recall	F-Score	Precision	Recall	F-Score
2 Tags	0.9695	0.9679	0.9687	0.9698	0.9683	0.9690
	±0.0040	±0.0056	±0.0046	±0.0035	±0.0048	±0.0040
3 Tags	0.9693	0.9686	0.9689	0.9703	0.9681	0.9692
	±0.0038	±0.0055	±0.0044	±0.0034	±0.0048	±0.0039
4 Tags	0.9694	0.9692	0.9693	0.9702	0.9676	0.9689
	±0.0038	±0.0053	±0.0043	±0.0034	±0.0048	±0.0040
5 Tags	0.9693	0.9692	0.9692	0.9703	0.9672	0.9687
	±0.0038	±0.0053	±0.0043	±0.0034	±0.0048	±0.0039

Table 4 Word segmentation performance (with standard errors) of the MM and CRF methods.

Method	Precision	Recall	F-Score
MM (MLC)	0.9881	0.7232	0.8351
	±0.0006	±0.0031	±0.0022
MM (MLC+BTEC1)	0.9093	0.9367	0.9228
	±0.0037	±0.0040	±0.0032
MM (BTEC1 Closed)	0.9106	0.7872	0.8444
	±0.0034	±0.0029	±0.0030
MM (BTEC1 Open)	0.9363	0.96243	0.7490
	±0.0074	±0.0013	±0.0020
CRF Character (5 tags)	0.9693	0.9692	0.9692
	±0.0038	±0.0053	±0.0043
CRF Syllable (5 tags)	0.9703	0.9672	0.9687
	±0.0034	±0.0048	±0.0039

Table.4 shows the performance of the CRF methods relative to the maximum matching baselines. It can be seen that the CRF models substantially outperform the MM systems in terms of the overall F-score, but the MM (MLC) method has a very high level of precision. The MM (BTEC1 Open) experiment used the same training and test data as the CRF model, and shows the in-domain performance using a small dictionary (approximately 8,500 entries). The MM (BTEC1 Closed) experiment used a dictionary from the entire 50,000-word training set that included the test data.

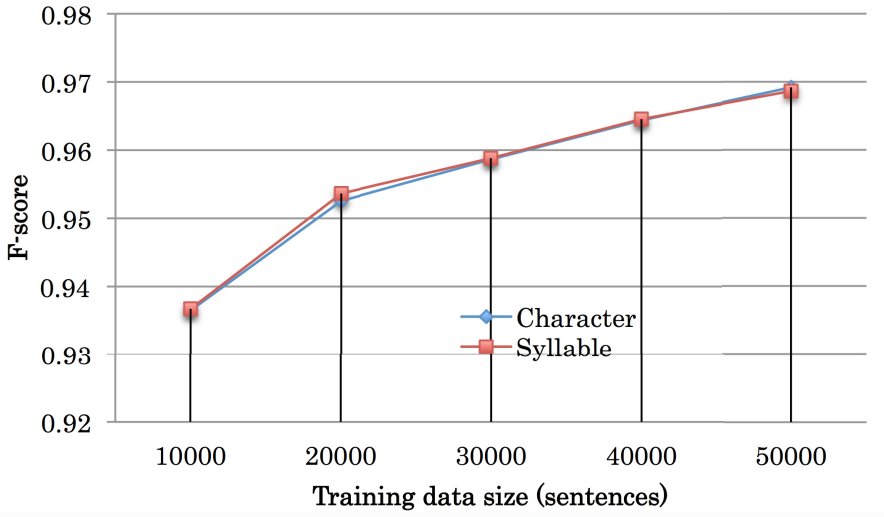


Fig. 4 F-Scores from training with CRF models on varying data set sizes.

In order to study how the CRF models behave with varying amounts of training data, we run a sequence of experiments that trained CRF models from 10K, 20K, 30K, 40K and 50K sentences respectively. From the results in Fig. 4, it is clear that the CRF model performs almost identically on character segmented and on syllable segmented data. Furthermore, the results show that the performance of the system is strongly linked to the data set size.

5 Error Analysis

Table.5 shows the most frequent labeling errors made by the CRF model when labeling syllable-segmented input. A list of syllables is shown for each error which represents the all of the syllables that gave rise to the error, listed in order of frequency.

It can be seen from the table that most of the errors in the top portion of the table are caused by syllables which end in one of the following vowels: ော, ါ, ိ, ီ, ား. Related errors occurred at the character level where the top 2 errors were: ိ, ား. Interestingly a large number of errors occurred on the following syllables: ပါ, မှာ, တာ, ထား, which often signify the ends of words, but may also occur quite frequently within words especially compound words. Single character consonants: မ, ရ, အ, ဝ, were also responsible for many errors. These mostly occur at the beginning of words labeled with '<', but can also occur at the beginning of words with the '+' label, and these cases appear to be difficult to disambiguate.

Table 5 Statistics on the most frequent 300 labeling errors from 10 experiments for syllable tagging together with all the associated syllables.

Reference	Output	Error Syllable	Percentage
|	+	ပေး,ပါ,တာ,သွား,ရ,ထား,ဘယ်,နေ,ပြန်,တစ်,တွေ,ရှိ,လိုက်,နား,တ,ဒီ,က	19.59
+	|	ပေး,သွား,ရှိ,တာ,ခေါ,တစ်,နစ်,တွေ,လာ,မှာ,လုပ်,ပြန်,ဆယ်,ပါ,ထား,တို,နေ	16.78
<	|	သွား,လို,ရှိ,ရ,ပေး,ဖြစ်,ထင်,နှစ်,လုပ်,တစ်,ထား	11.44
>	+	ပါ,ရ,မ,နာ,အ,နေ,စ,မှာ	9.49
|	<	ရ,ပါ,ချင်,နေ,လို,ဖြစ်,ပေး	8.54
+	>	မ,ပါ,သ,စ,အ	8.29
<	+	အ,မ,ဘယ်,ဆောင်,နေ,ဒီ,ကြ	7.96
-	<	ရ,ပါ,နိုင်,ပေး	6.18
+	<	အ,ဘယ်,မ,နေ,နည်း	5.66
<	-	ပါ,နိုင်,ရ	3.38
>	|	လောက်,ရွက်,ရာ,တာ,တွေ,နား	1.62
|	>	လောက်,ပါ,နည်း	0.95

6 Conclusion

In this paper we have studied the application of CRF models to perform supervised word identification of Myanmar text. The performance of the CRF models was compared against a baseline model based on maximum matching that is close to the current state-of-the-art in Myanmar word segmentation. Our results show that the overall performance of the CRF models, measured in terms of F-score was substantially higher than the maximum matching baseline. We were also able to show that the CRF model was able to perform word segmentation equally well from either Myanmar characters or syllables. Experiments on data set size revealed that the CRF is still improving even on the largest training data set size of 50,000 sentences, and therefore we believe that the acquisition of more data is critically important in improving the segmentation accuracy of the system. In future work, we intend to increase the size of the manually segmented corpus since our experiments indicated that this was likely to deliver significant improvement in performance.

Acknowledgments We thank Ms. Aye Mya Hlaing, Ms. Aye Myat Mon, Ms. Hay Mar Soe Naing, Ms. Min Min Oo, Ms. Nyo Lay Myint, Ms. Sann Su Su Yee and Ms. Thae Nu Htwe from the University of Computer Studies, Yangon (UCSY), Myanmar for their help in segmenting 50,000 sentences of BTEC1 Myanmar Corpus manually. We also thanks Dr. Chenchen Ding (Multilingual Translation Lab., NICT) for his python script tagger (character level).

References

1. Pa, W.P., Thein, N.L.: Myanmar Word Segmentation using Hybrid Approach. In: Proceedings of 6th International Conference on Computer Applications, Yangon, Myanmar, pp. 166–170 (2008)

2. Kikui, G., Yamamoto, S., Takezawa, T., Sumita, E.: Comparative study on corpora for speech translation. IEEE Transactions on Audio, Speech, and Language Processing, 14(5), 1674–1682 (2006)
3. Thu, Y.K., Finch, A., Sagisaka, Y., Sumita, E.: A Study of Myanmar Word Segmentation Schemes for Statistical Machine Translation. In: Proceedings of 12th International Conference on Computer Applications, Yangon, Myanmar, pp. 167–179 (2014)
4. Lafferty, J., McCallum, A., Pereira, F.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: Proceedings of the 18th International Conf. on Machine Learning, pp. 282–289 (2001)
5. Thet, T.T., Na, J.-C., Ko, W.K.: Word Segmentation for the Myanmar language. Journal of Information Science **34**(5), 688–704 (2008)
6. Htay, H.H., Murthy, K.N.: Myanmar Word Segmentation Using Syllable Level Longst Matching, the 6th Workshop on Asian Language. Resources **2008**, 41–48 (2008)
7. Liu, Y., Tan, Q., Shen, K.X.: The Word Segmentation Methods for Chinese Information Processing. Quing Hua University Press and Guang Xi Science and Technology Press, 36 (1994) (in Chinese)
8. Myanmar English Dictionary, Myanmar Language Commission, Myanmar, 2012 Edition
9. Myanmar Grammar, Myanmar Language Commission, Myanmar (2000)
10. Taku Kudo: CRF++ An open source toolkit for CRF (2005). http://crfpp.sourceforge.net/

Index Structure for Nearest Neighbors Search with Required Keywords on Spatial Database

Su Nandar Aung and Myint Myint Sein

Abstract There is more and more commercial and research interest in nearest neighbor objects for location-based search from spatial database. Specially, a spatial keyword query takes a user location and user-supplied keywords as arguments and returns objects that is nearest k objects from user current location and textually relevant to the user required keyword. In these systems, user can type one or more word for required keyword. To find the result that contains at least one required keyword, it is important to support Boolean OR semantic keyword search on spatial database. In this paper we study how to answer such queries efficiently for both kNN query and range query. This paper proposes new index structure that combines K-d tree and inverted file. K-d tree efficiently support for both nearest neighbor and range queries. We also discuss Boolean OR Semantic keyword search for the user's required keywords.

Keywords Spatial keyword queries · Problem statement · Proposed index · K-NN keyword search algorithm · Range keyword search algorithm

1 Introduction

Spatial database systems manage large collections of spatial data, which apart from spatial attributes contain non spatial information. Spatial data are data that have a location (spatial) and mainly required for Geographic Information Systems (GIS) whose information is related to geographic locations. GIS model supports spatial data types, such as point, line and polygon. A geospatial collections increase in size, the demand of efficient processing of spatial queries with text constraints becomes more prevalent.

Due to the popularity of keyword search, particularly on the Internet, many of these applications allow the user to provide a set of keywords that the spatial

S.N. Aung(✉) · M.M. Sein(✉)
University of Computer Studies, Yangon, Myanmar
e-mail: {sunandaraung,myint}@ucsy.edu.mm

T.T. Zin et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 388,
DOI: 10.1007/978-3-319-23207-2_47

objects should contain, in their name or description or categories. Spatial keyword search is an important tool in exploring useful information from a spatial database and has been studied for years. The query consists of a spatial location, a set of keywords and a parameter k and the answer is a list of objects ranked according to a combination of their distance to the query point and the relevance of their text description to the query keyword. *The spatial relevance* is measured by the distance between the location associated with the candidate document to the query location, and *the textual relevance* is said to be textually relevant to a query if object contains queried keywords.

During the design of a spatial index, issues that need to be minimized are:

(a) The area of covering rectangles maintained in internal nodes,
(b) The overlaps between covering rectangles for indexes developed based on the overlapping native space indexing approach,
(c) The number of objects being duplicated for indexes developed based on the non-overlapping native space indexing approach
(d) The directory size and its height.
(e) The index shows effective storage utilization.
(f) The index answers queries efficiently.
(g) The index can answers possible result with user's required keyword within minimum time.

Many index structures that have been proposed in recent years mainly use R-tree and then combine with inverted file, namely the families of IR-tree [4, 5, 6, 7, 8, 9, and 10]. All use R-tree for spatial (latitude/longitude) index and inverted file for textual index. They all created hybrid index structure according to the combination schemes: (1) *Text first loose combination scheme*, employs the inverted as the top-level index and then arrange the postings in each inverted list in a spatial structure. (2) *Spatial-first loose combination scheme* employs the spatial index as the top-level index and its leaf nodes contain inverted files or bitmaps for the text information of objects contained in the nodes. (3) *Tight combination index*combines a spatial and a text index tightly such that both types of information can be used to prune the search space simultaneously during query processing.

The construction of an efficient index structure should take into account overlaps between nodes and coverage of a node. Minimization of a node coverage leads to more precise searching within the tree and minimization of the overlap between nodes reduces the number of paths tested in the tree during a search that can reduce search time. As the data objects in the R-tree can be overlapping and covering each other, the search process in the R-tree might suffer from unnecessary node visits and higher IO cost [16]. Moreover, the IR-trees suffer from high update cost. Each node has to maintain an inverted index for all the keywords of documents associated with this node's MBR. When a node is full and split into two new nodes, all the textual information in the node has to be re-organized [1]. As the R-tree need to reorganized, it suffers from higher CUP costs. This paper intends to reduce IO costs, CUP costs and searching time for kNN keyword search. In this paper, we study Boolean OR semantic for spatial keyword queries.

The Boolean OR semantics allow the query keywords to partially match the data in the database. This is a more general case.

For example, queries like "Japanese Sushi Restaurant", a user does not have all the complete keywords, but "Japanese" or "Sushi" or "Restaurant" can be recommended if they are closed to the user's current location for kNN query and if they are within user's defined range. More candidates will be examined in the query processing of OR semantics.

This paper includes the following contributions:

1) The main contribution is to create index structure that combine K-d tree and inverted file for efficiently process spatial keyword queries.
2) Nearest neighbor keyword search algorithm and range algorithm is developed using the proposed index structure to efficiently answer spatial keyword query using OR semantic and to explore useful information that user required.

2 Related Works

There has been lot of interest in building geographic information retrieval system. Spatial Keyword search has been well studied for years due to its importance to commercial search engines. Various types of spatial keyword queries have been proposed. For spatial keyword search, the index structure is created for both spatial and textual relevance. Most index structures [5],[6],[8],[9],[10] use R-tree and its variants as spatial index and inverted file for text index.

They all combine both indices depending on the combination schemes [15]. Among them [8] integrates signature file instead of inverted file into each node of the R-tree. Inverted file-R*tree (IF-R*) and R*-tree-inverted file (R*-IF) [10] are two geo-textual indices that loosely combine the R*-tree and inverted file. Hariharan et al. R. G¨obel, A. Henrich, R. Niemann, and D. Blank [8] proposed the KR*-tree. This paper proposed a framework for GIR systems and focus on indexing strategies.I. D. Felipe, V. Hristidis, and N. Rishe [9] uses R*-tree for spatial index and inverted file for text index. Cary et al, [5] proposed SKI that combines and R-tree with an inverted index by the inclusion of spatial references in posting lists.

In [5] the posting list of term contains all its term bitmaps rather than documents. The IR tree [6] creates each nodes of the R-tree with a summary of the text content of the objects in the corresponding subtree. Li et al. proposed an index structure, which is also called IR tree that stores one integrated inverted file for all the nodes. X. Cao, L. Chen, G. Cong, C. S. Jensen, Q. Qu, A. Skovsgaard, D. Wu, and M. L. Yiu [3] proposed S2I index structure based on R-tree and inverted file. The objects in [3] are stored differently according to the document frequency and infrequency of the term.

D. Zhang, K.L. Tan, Anthony K.H. Tung[1] proposed I^3 (Integrated Inverted Index), which adopts the Quad tree structures to hierarchically partition the data space into cells. The basis unit of I^3 is the keyword cell, which captures the spatial locality of a keyword. X.Cao, G.Cong, Christian S. Jensen, Jun.J. Ng, BengC.Ooi, N.T. Phan, D. Wu [15] proposes a Web Object Retrieval System (SWORS) that is

capable of efficiently retrieving spatial web objects that satisfy spatial keyword queries. This system use IR tree and inverted file for index. It supports two types of queries that are location aware top-k text retrieval (Lkt) query and spatial keyword group query.

3 Problem Statement

Let D is a spatial database that contains D= $\{o_1,o_2,o_3\ldots..,o_n\}$ such that every object o in D has many attributes $<o_{id}, o_l, o_d>$ where o_{id} is an identifier of an object, o_l is a spatial location that contain latitude and longitude and o_d is an text document of each object for keyword querying.A *keyword query* $q_k$is a set of keywords $k_1, k_2, k_3,\ldots., k_m$. The result is a set of objects ordered by the relevance of their textual description to the query keywords.

Boolean Range Keyword Query: Let D is a spatial database that contains D= $\{o_1,o_2,o_3\ldots,o_n\}$ such that every object o in D has many attributes $<o_{id}, o_l, o_d>$ where o_{id} is an identifier of an object, $o_l$is a spatial location that contain latitude and longitude and o_d is an text document of each object for keyword querying.

Let $q=<q_k, q_r>$ be a range keyword query where q_k is user required keywords w1…..wm and q_r is the user desired range (km). A query q return all objects in D that contain at least one similar keywords q_k={w1, w2,…., wm} and belong to the range qr .

$$Ans(q) = \begin{cases} o_l \ within\ given\ range\ q_r \\ \quad \vdots \exists_w \in q_k,\ \ w \in o_d \end{cases}$$

Boolean kNN Keyword Queries: Let $q=<q_k, q_l, k>$ be a Boolean kNN query where q_k is user required keywords $w_1\ldots..w_m$, q_l is a user current location (latitude, longitude) and k is the number of result objects. A query q return k objects o_k from D that are nearest neighbor of $q_l$with the highest scores $\alpha(p, ql), \alpha(p_1, ql) \geq \alpha(p_2, ql) \geq \alpha(p_3, ql) \geq \cdots \geq \alpha(p_4, ql),$ according to the Euclidean Distance and Boolean Model in which corresponding point contain at least one required keywordsq_k={w_1, w_2,…., w_m}.

$$Ans(q) = \begin{cases} o_l first\ ordered\ by\ distance\ (o_l, q_l) \\ \quad \vdots \exists_w \in q_k,\ \ w \in o_d \end{cases}$$

4 Proposed System

In this proposed system K-d tree loosely combined with inverted file. K-d tree is used for spatial queries and inverted file is used for keywords information that is the most efficient index for text information retrieval.For each node of K-d tree, an inverted file is created for indexing the text components of objects contained in the node. As K-d trees represent a disjoint partition, the proposed system can't cause more IO costs and also K-d trees don't need to rebalance the textual information so the proposed can reduce update cost (CPU Costs).

Most geo-textual indices use the inverted file for text indexing. The frequency information is not included in the inverted file that is developed to handle Boolean queries. Inverted file can be used to check the query keywords contain or not. K-d tree structure is known as point indexing structures as it is designed to index data objects which are points in a multi-dimensional space. It can be used efficiently for nearest neighbor query and range query. This paper proposes nearest neighbor keyword search algorithm and range keyword search algorithm using K-d tree and inverted file.

Table 1 Example Dataset

id	Latitude	Longitude	Keywords
Obj1	16.779568	96.152687	Mobile, Shopping, Mall, Telecommunication, Electronics, Tools
Obj2	16.779533	96.15269	May, Shopping, Center, Super Market
Obj3	16.813517	96.08475	Cat, Walk, Foot, Wear
Obj4	16.779565	96.135581	NorthPoint, Shopping, Center, Super, Market, Food, Drink
Obj5	16.881351	96.152549	Gamonpwint, Shopping, Center, Super, Market, Food, Drink
Obj6	16.779581	96.169647	Gamonpwint, Shopping, Center, Super, Market, Food, Drink
Obj7	16.779568	96.152719	Asia, Shopping, Center, Super, Market
Obj8	16.830324	96.186432	Moon, Bakery, Food, Drink

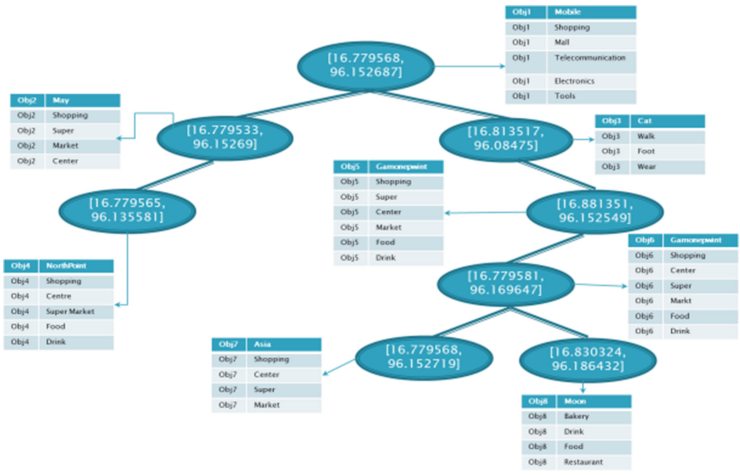


Fig. 1 Proposed Index Structure for Dataset of Table 1

5 K-NN Keyword Search Algorithm

The algorithm-1 returns the closest points to a given user's current location according to a certain distance function. When the algorithm explores some points of the kd-tree, it starts computing the distance between this points and query point and then check at least one of the required keywords contain using Boolean Semantic OR model.

Algorithm1. K-NN Keyword Search in Proposed Index Structure

```
NNKeywordSearch (T,Q)
     T: kd tree;
Q: Query that contains current location Q.l, required keyword Q.key, number of required nearest neighbours
    objects Q.k;
L: Arraylist;
pqResult: Priority Queue;
count ← 0;
pqResult ← NNSearch (T, Q);
while count <Q.k do
    L.add (pqResult.remove());
return L;

NNSearch (T, Q)
pq: Proirity Queue;
pqResult: Priority Queue;
Search: tuple kd tree, bounding_box, potential_distance and tuple;
nnPoint: undefine;
minDistance: infinity ( ∞)
pq.add( Search ( T, B_Box, 0));
whilepq.size>0 and pq.TOP(().potential_distance<minDistance do
        T ← pq.TOP().kdtree;
B_Box ← pq.TOP ().bounding_box ;
pq.remove();
    if T≠ leaf then
      point ← T.key;
      i ← T.discr;
      distance ← DISTANCE ( point.l , Q.l);
     if distance<minDistance&&∑_{j=1}^{qk.count} keyword_j ∈ ∑_{i=1}^{n} keyword_i
     pqResult.ADD(point);
     minDistance ← distance;
BOUNDINGBOX (left_BB,right_BB,BB,point[i])
potential_distance ←MINDISTANCE(left_BB,Q.l);
If potential_distance< distance then
    pq.ADD(Search(T.left, left_BB, potential_distance));
potential_distance← MINDISTANCE (right_BB,Q.l);
If potential_distance< distance   then
        pq.ADD(Search(T.right, right_BB,potential_distance));
returnpqResult;
```

In the algorithm-1, the procedure ComputeBoundingBoxes(…) returns the bounding boxes lBB and rBB for the left and the right subtrees, respectively. The function MinimumDistance(BB; c) returns the potential distance between any point located inside the bounding box BB and query point. The DISTANCE (…) procedure calculates the distance between two points using Euclidean distance. i.e;

$$d(q, p) = \sqrt{(q_{lat} - p_{lat})^2 + (q_{lon} - p_{lon})^2}$$

6 Range Keyword Search Algorithm

Algorithm-2 is a proposed range keyword search procedure. The procedure RANGEKESEARCH returns all points 'p' such that d (q,p) ≤ r and $\sum_{j=1}^{qk.count} keyword \in \sum_{i=1}^{n} p.word_i$. Token each word from user input keywords and then saves in array qk. Use Boolean OR semantics model to check at least one required keywords contain or not in inverted file of each point such that,

$$Ans = \begin{cases} 1, & if\ one\ or\ more\ \ keywords\ contain\ in\ p.word \\ o\ , otherwise \end{cases}$$

Algorithm2. Range Keyword Search Algorithm Using Proposed Index Structure

Input: user's required keyword, K-d tree, query point, range, Max/Min BB
pq: priority queue
qk : array
RANGEKESEARCH (keyword,T,BB,q,r)
if T=leaf then return
p<–T.key; i<–T.discr;
distance<– DISTANCE(q,p);
if distance ≤ r and $\sum_{j=1}^{qk.count} keyword \in \sum_{i=1}^{n} p.word_i$ then
pq.PUSH (p,distance);
COMPUTEBGBOXES (lBB,rBB,p[i],i)
if INTERSECTS (lBB,q,r) then
RANGEKESEARCH (keyword, T.left, lBB, c, radius)
if INTERSECTS (rBB,q,r) then
RANGEKESEARCH(keyword, T.right, rBB, c, radius)

The procedure COMPUTEBBOXES( ) calculates the bounding boxes lBB and rBB for the left and for the right sub tree, respectively. The procedure INTERSECTS (…) tells if the bounding box BB intersects with the region that satisfies the distance constraints. If the intersection is non-empty, the sub tree to be explored.The DISTANCE (…) procedure calculates the distance between two points using Euclidean distance $d(q,p) = \sqrt{(q_{lat} - p_{lat})^2 + (q_{lon} - p_{lon})^2}$.

7 Architecture of Proposed System

The propose system adopts the browser-server model for desktop and laptop computer. Figure 2 shows the user interface for the proposed system. Users can input their queries through the web browser and the queries are sent to the server for processing. After the queries are processed, the results are sent back and displayed using Google Maps in the users' browser. Queries are sent from the browser to the server by the HTTP post operation. The browser side use Google Map API to provide interfaces to users for generating queries and viewing the returned spatial web objects. Users can specify the current's location by clicking a location in Google Map to get the latitude and longitude of that location and can type the required keywords.

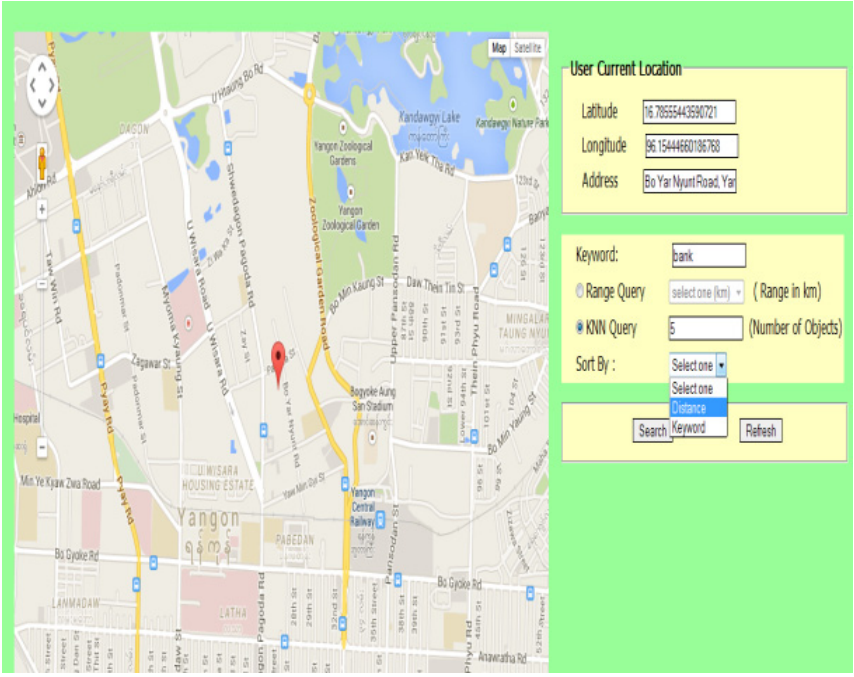


Fig. 2 User Interface for the Proposed System

Then user has to choose range query or kNN query. If user choose range query, user must choose desired range (km). Otherwise, user must be input the required number of objects k. Finally user needs to choose sort by type. The query is sent to the server and then the results are sorted by the distance or keyword and then are displayed on Google Maps in the browser.

8 Experimental Results

Fig. 3 shows the index construction time (second) depending on the size of data-sets. Fig. 4 compare the searching time (second) depending on the number of required keywords between using proposed index structure and other index that combine R-tree and inverted file in kNN query. Fig. 5 shows the searching time depending on the varying number of objects k in kNN query. Fig. 6 shows the searching time in range keyword query and Fig. 7 shows the searching time depend on the number of required keyword.

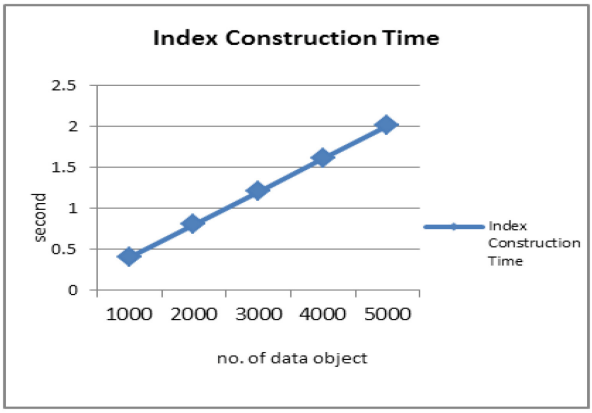


Fig. 3 Index Construction Time

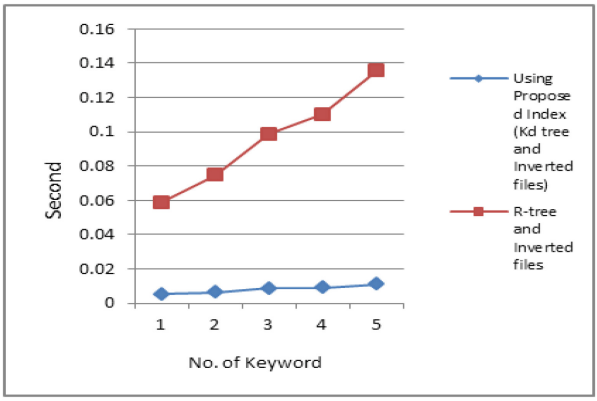


Fig. 4 Searching Time for varying number of keywords in kNN query

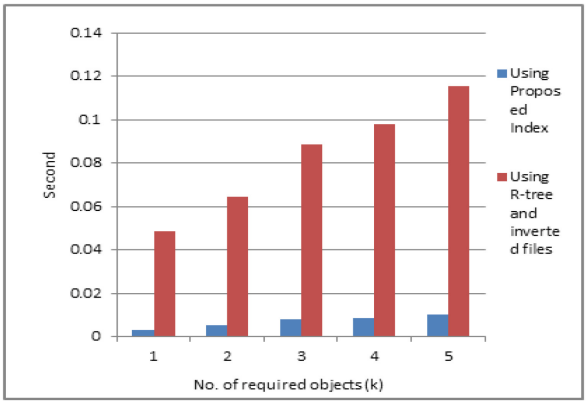


Fig. 5 Searching Time for varying number of required objects in kNN query

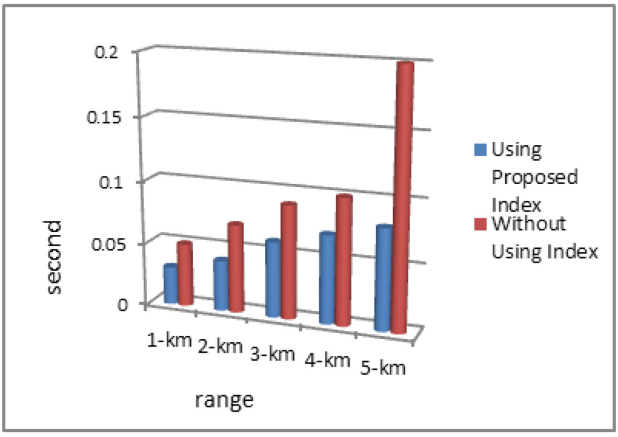


Fig. 6 Searching Time in range keyword search

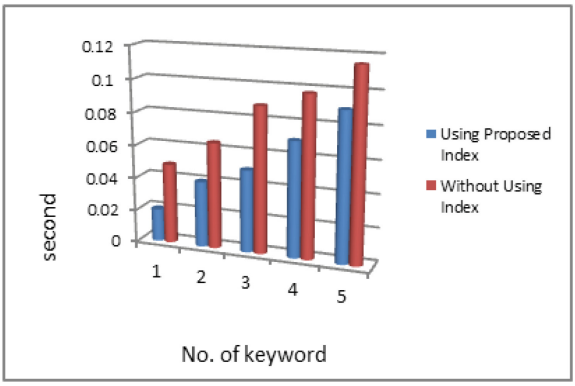


Fig. 7 Searching Time for varying number of required keyword in range query

9 Conclusion and Further Extension

This paper presented hybrid index structure for range keyword query and kNN keyword query searching with minimum IO costs and CPU costs. This index structure can avoid searching in overlapping area. So it can reduce searching time in overlap area. Moreover, it can't cause node overflow, so it doesn't need to re-organize the textual data and spatial data. Many Further extensions can be considered for efficient hybrid index structure for spatial database. As a further extension, we'll add an efficient spatial approximate keyword search and Boolean AND Semantics keyword search within given range and nearest neighbour. Furthermore, we'll consider Approximate Nearest Neighbors search with required keyword by using the proposed index structure.

References

1. Zhang, D., Tan, K.L., Tung, A.K.H.: Scalable top-k spatial keyword search. In: EDBT/ICDT 2013, March 18–22, 2013
2. Chen, L., Cong, G., Jensen, C.S., Wu, D.: Spatial keyword query processing: an experimental evaluation In: Proceddings of the VLDB Endowment, vol.6, no.3 (2013)
3. Cong, G., et al.: Spatial keyword queries. In: Yu, H., Yu, G., Hsu, W., Moon, Y.-S., Unland, R., Yoo, J. (eds.) DASFAA Workshops 2012. LNCS, vol. 7240, p. 250. Springer, Heidelberg (2012)
4. Rocha-Junior, J.B., Gkorgkas, O., Jonassen, S., Nørvåg, K.: Efficient processing of top-k spatial keyword queries. In: Pfoser, D., Tao, Y., Mouratidis, K., Nascimento, M.A., Mokbel, M., Shekhar, S., Huang, Y. (eds.) SSTD 2011. LNCS, vol. 6849, pp. 205–222. Springer, Heidelberg (2011)
5. Li, Z., Lee, K.C.K., Zheng, B., Lee, W.-C., Lee, D.L., Wang, X.: Ir-tree: An efficient index for geographic document search. IEEE TKDE **23**(4), 585–599 (2011)
6. Cary, A., Wolfson, O., Rishe, N.: Efficient and scalable method for processing top-k spatial boolean queries. In: Gertz, M., Ludäscher, B. (eds.) SSDBM 2010. LNCS, vol. 6187, pp. 87–95. Springer, Heidelberg (2010)
7. Cong, G., Jensen, C.S., Wu, D.: Efficient retrieval of the top-k most relevant spatial web objects. PVLDB **2**(1), 337–348 (2009)
8. Göbel, R., Henrich, A., Niemann, R., Blank, D.: A hybrid index structure for geo-textual searches. In: CIKM, pp. 1625–1628 (2009)
9. Felipe, I.D., Hristidis, V., Rishe, N.: Keyword search on spatial databases. In: ICDE, pp. 656–665 (2008)
10. Hariharan, R., Hore, B., Li, C., Mehrotra, S.: Processing spatial-keyword (sk) queries in geographic information retrieval (gir) systems. In: SSDBM, pp. 16 (2007)
11. Zhou, Y., Xie, X., Wang, C., Gong, Y., Ma, W.-Y.: Hybrid index structures for location-based web search. In: CIKM, pp. 155–162 (2005)
12. Kakde, H.M.: Range Searching using Kd Tree (2005)
13. Guttman, A.: R-trees: a dynamic index structure for spatial searching. In: SIGMOD, pp. 47–57 (1984)
14. Ooi, B.C., Sacks-Davis, R., Han, J.: Indexing in Spatial Databases
15. Cao, X., Cong, G., Christian, S.J., Ng, J.J., Ooi, B.C., Phan, N.T., Wu, D.: SWROS: A System for the Efficient Retrieval of Relevant Spatial Web Objects
16. Theodoridis, Y., Sellis, T.: Optimization Issues in R-tree Construction. Technical Report KDBSLAB-TR-93-08

Author Index