

Jerry Chun-Wei Lin
Jeng-Shyang Pan
Shu-Chuan Chu
Chien-Ming Chen *Editors*

Genetic and Evolutionary Computing

Proceedings of the Eleventh
International Conference on Genetic
and Evolutionary Computing,
November 6–8, 2017,
Kaohsiung, Taiwan

Advances in Intelligent Systems and Computing

Volume 579

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.

Advisory Board

Chairman

Nikhil R. Pal, Indian Statistical Institute, Kolkata, India

e-mail: nikhil@isical.ac.in

Members

Rafael Bello Perez, Universidad Central “Marta Abreu” de Las Villas, Santa Clara, Cuba

e-mail: rbellop@uclv.edu.cu

Emilio S. Corchado, University of Salamanca, Salamanca, Spain

e-mail: escorchado@usal.es

Hani Hagrass, University of Essex, Colchester, UK

e-mail: hani@essex.ac.uk

László T. Kóczy, Széchenyi István University, Győr, Hungary

e-mail: koczy@sze.hu

Vladik Kreinovich, University of Texas at El Paso, El Paso, USA

e-mail: vladik@utep.edu

Chin-Teng Lin, National Chiao Tung University, Hsinchu, Taiwan

e-mail: ctlin@mail.nctu.edu.tw

Jie Lu, University of Technology, Sydney, Australia

e-mail: Jie.Lu@uts.edu.au

Patricia Melin, Tijuana Institute of Technology, Tijuana, Mexico

e-mail: epmelin@hafsamx.org

Nadia Nedjah, State University of Rio de Janeiro, Rio de Janeiro, Brazil

e-mail: nadia@eng.uerj.br

Ngoc Thanh Nguyen, Wroclaw University of Technology, Wroclaw, Poland

e-mail: Ngoc-Thanh.Nguyen@pwr.edu.pl

Jun Wang, The Chinese University of Hong Kong, Shatin, Hong Kong

e-mail: jwang@mae.cuhk.edu.hk

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

Jerry Chun-Wei Lin · Jeng-Shyang Pan
Shu-Chuan Chu · Chien-Ming Chen
Editors

Genetic and Evolutionary Computing

Proceedings of the Eleventh International
Conference on Genetic and
Evolutionary Computing, November 6–8, 2017,
Kaohsiung, Taiwan

Editors

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

Jeng-Shyang Pan
Fujian Provincial Key Laboratory of Big
Data Mining and Applications
Fujian University of Technology
Fuzhou, Fujian
China

Shu-Chuan Chu
School of Computer Science,
Engineering and Mathematics
Flinders University
Bedford Park, SA
Australia

Chien-Ming Chen
School of Computer Science
and Technology
Harbin Institute of Technology Shenzhen
Graduate School
Shenzhen, Guangdong
China

ISSN 2194-5357

ISSN 2194-5365 (electronic)

Advances in Intelligent Systems and Computing

ISBN 978-981-10-6486-9

ISBN 978-981-10-6487-6 (eBook)

<https://doi.org/10.1007/978-981-10-6487-6>

Library of Congress Control Number: 2017956061

© Springer Nature Singapore Pte Ltd. 2018

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by Springer Nature

The registered company is Springer Nature Singapore Pte Ltd.

The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

This volume composes the proceedings of the Eleventh International Conference on Genetic and Evolutionary Computing (ICGEC 2017), which is hosted by Fujian University of Technology and is held in Kaohsiung, Taiwan, on November, 6–8, 2017. ICGEC 2017 is technically co-sponsored by Springer, Fujian Provincial Key Laboratory of Big Data Mining and Applications, Fujian University of Technology, National University of Kaohsiung, Harbin Institute of Technology, National Kaohsiung University of Applied Science, and VSB-Technical University of Ostrava. It aims to bring together researchers, engineers, and policymakers to discuss the related techniques, to exchange research ideas, and to make friends.

Twenty-four excellent papers were accepted for the final proceeding. 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 Fujian University of Technology, China, in making ICGEC 2017 possible and also appreciate the great help from National Kaohsiung University of Applied Science for local organizing the conference.

August 2017

Jerry Chun-Wei Lin
Jeng-Shyang Pan
Shu-Chuan Chu
Chien-Ming Chen

Organization

Organizing Committee

Honorary Chairs

Xin Tong	Fujian University of Technology, China
Jeng-Shyang Pan	Fujian University of Technology, China
Leon Shyue-Liang Wang	National University of Kaohsiung, Taiwan

Advisory Committee Chairs

Tzung-Pei Hong	National University of Kaohsiung, Taiwan
Ponnuthurai Nagaratnam Suganthan	Nanyang Technological University, Singapore
Philip S. Yu	University of Illinois at Chicago, USA

Conference Chairs

Jerry Chun-Wei Lin	Harbin Institute of Technology Shenzhen Graduate School, China
Vaclav Snasel	VŠB-Technical University of Ostrava, Czech Republic
Pyke Tin	University of Computer Studies, Myanmar

Program Committee Chairs

Philippe Fournier-Viger	Harbin Institute of Technology Shenzhen Graduate School, China
I-Hsien Ting	National University of Kaohsiung, Taiwan

Local Organization Chairs

Mong-Fong Horng	National Kaohsiung University of Applied Sciences, Taiwan
Tsu-Yang Wu	Fujian University of Technology, China

Electronic Media Chairs

Pei-Wei Tsai	Swinburne University of Technology, Australia
Thi-Thi Zin	University of Miyazaki, Japan

Publication Chairs

Chien-Ming Chen	Harbin Institute of Technology Shenzhen Graduate School, China
Bay Vo	Ho Chi Minh City University of Technology, Vietnam
Jimmy Ming-Tai Wu	Harbin Institute of Technology Shenzhen Graduate School, China

Finance Chair

Jui-Fang Chang	National Kaohsiung University of Applied Sciences, Taiwan
----------------	--------------------------------------------------------------

Program Committee

Chien-Ming Chen	Harbin Institute of Technology Shenzhen Graduate School, China
George Chang	Kean University, USA
Philippe Fournier-Viger	Harbin Institute of Technology Shenzhen Graduate School, China
Vicente García Díaz	University of Oviedo, Spain

Wensheng Gan	Harbin Institute of Technology Shenzhen Graduate School, China
Tzung-Pei Hong	National University of Kaohsiung, Taiwan
Jean Hok Yin Lai	Hong Kong Baptist University, Hong Kong
Jerry Chun-Wei Lin	Harbin Institute of Technology Shenzhen Graduate School, China
Eric Hsueh-Chan Lu	National Cheng Kung University, Taiwan
Ivan Lee	University of South Australia, Australia
Wen-Yang Lin	National University of Kaohsiung, Taiwan
Kawuu W. Lin	National Kaohsiung University of Applied Sciences, Taiwan
Ashish Sureka	ABB Corporate Research Center, India
Ja-Hwung Su	Cheng Shiu University, Taiwan
I-Hsien Ting	National University of Kaohsiung, Taiwan
Pei-Wei Tsai	Swinburne University of Technology, Australia
Chun-Wei Tsai	National Chung Hsing University, Taiwan
Miroslav Voznak	VŠB-Technical University of Ostrava, Czech Republic
Bay Vo	Ho Chi Minh City University of Technology, Vietnam
Tsu-Yang Wu	Fujian University of Technology, China
Jimmy Ming-Tai Wu	Harbin Institute of Technology Shenzhen Graduate School, China
Cheng-Wei Wu	National Chiao Tung University, Taiwan
Mu-En Wu	SooChow University, Taiwan
Unil Yun	Sejong University, Korea
Ji Zhang	University of Southern Queensland, Australia
Thi Thi Zin	University of Miyazaki, Japan

Contents

Evolutionary Computation

A Many-Objective Evolutionary Algorithm with Reference Point-Based and Vector Angle-Based Selection	3
Chen-Yu Lee, Jia-Fong Yeh, and Tsung-Che Chiang	
Freeway Travel Time Prediction by Using the GA-Based Hammerstein Recurrent Neural Network	12
Ru-Kam Lee, Yi-Che Yang, Jun-Hong Chen, and Yi-Chung Chen	
A PIP-Based Approach for Optimizing a Group Stock Portfolio by Grouping Genetic Algorithm	20
Chun-Hao Chen and Chih-Hung Yu	
A Novel Genetic Algorithm for Resource Allocation Optimization in Device-to-Device Communications	26
Yung-Fa Huang, Tan-Hsu Tan, and Bor-An Chen	

Data Mining and Its Applications

Mining Erasable Itemsets Using Bitmap Representation	37
Wei-Ming Huang, Tzung-Pei Hong, Guo-Cheng Lan, Ming-Chao Chiang, and Jerry Chun-Wei Lin	
Identifying Suspicious Cases in the Hong Kong Stock Market Using Commentators' Stock News	44
Li Quan and Jean Lai	
A New Conceptual Model for Big Data Analysis	52
Thi Thi Zin, Pyke Tin, and Hiromitsu Hama	
An Hybrid Multi-Core/GPU-Based Mimetic Algorithm for Big Association Rule Mining	59
Youcef Djenouri, Asma Belhadi, Philippe Fournier-Viger, and Jerry Chun-Wei Lin	

Updating the Discovered High Average-Utility Patterns with Transaction Insertion 66
Tsu-Yang Wu, Jerry Chun-Wei Lin, Yinan Shao, Philippe Fournier-Viger, and Tzung-Pei Hong

Image and Multimedia Processing

Applying Image Processing Technology to Region Area Estimation 77
Yi-Nung Chung, Yun-Jhong Hu, Xian-Zhi Tsai, Chao-Hsing Hsu, and Chien-Wen Lai

Face Recognition under Lighting Variation Conditions Using Tan-Triggs Method and Local Intensity Area Descriptor 84
Chi-Kien Tran, Duc-Tinh Pham, Chin-Dar Tseng, and Tsair-Fwu Lee

Analysis of the Dynamic Co-purchase Network Based on Image Shape Feature 93
Xiaoyin Li and Jean Lai

VQ Compression Enhancer with Huffman Coding 101
Chin-Feng Lee, Chin-Chen Chang, and Qun-Feng Zeng

Adaptive Steganography Method Based on Two Tiers Pixel Value Differencing 109
Chi-Yao Weng, Yen-Chia Huang, Chin-Feng Lee, and Dong-Peng Lin

Intelligent Systems

A House Price Prediction for Integrated Web Service System of Taiwan Districts 121
Chia-Chen Fan, Shyan-Ming Yuan, Xuebai Zhang, and Yu-Chuan Lin

Commonsense-Knowledge Based Inference Engine 128
Zhengdao Peng and Jean Lai

Analysis of Users' Emotions Through Physiology 136
Bohdan Myroniv, Cheng-Wei Wu, Yi Ren, and Yu-Chee Tseng

Research on Temperature Rising Prediction of Distribution Transformer by Artificial Neural Networks 144
Wenxin Zhang, Jeng-Shyang Pan, and Yen-Ming Tseng

Development of Audio and Visual Attention Assessment System in Combination with Brain Wave Instrument: Apply to Children with Attention Deficit Hyperactivity Disorder 153
Chin-Ling Chen, Yung-Wen Tang, Yong-Feng Zhou, and Yue-Xun Chen

Decision Support Systems

Markov Queuing Theory Approach to Internet of Things Reliability 165
Thi Thi Zin, Pyke Tin, and Hiromitsu Hama

Some Characteristics of Nanyaseik Area Corundum and Other Assorted Gemstones in Myanmar 173
Htin Lynn Aung and Thi Thi Zin

Exploring Gemstones in Northern Part of Myanmar 182
Htin Lynn Aung and Thi Thi Zin

Encryption and Security

Attacks and Solutions of a Mutual Authentication with Anonymity for Roaming Service with Smart Cards in Wireless Communications 191
Tsu-Yang Wu, Bin Xiang, Guangjie Wang, Chien-Ming Chen, and Eric Ke Wang

Comments on Islam Et Al.’s Certificateless Designated Server Based Public Key Encryption with Keyword Search Scheme 199
Tsu-Yang Wu, Chao Meng, King-Hang Wang, Chien-Ming Chen, and Jeng-Shyang Pan

Author Index 207

Evolutionary Computation

A Many-Objective Evolutionary Algorithm with Reference Point-Based and Vector Angle-Based Selection

Chen-Yu Lee, Jia-Fong Yeh, and Tsung-Che Chiang^(✉)

Department of Computer Science and Information Engineering,
National Taiwan Normal University, Taipei, Taiwan (R.O.C.)
jerry789520@gmail.com, j530621@gmail.com
tcchiang@ieee.org

Abstract. In this paper we proposed a many-objective evolutionary algorithm by combining the reference point-based selection in NSGA-III and the vector angle-based selection in VaEA. Performance of the proposed algorithm is verified by testing on the negative version of four DTLZ functions. The proposed algorithm is better than NSGA-III and is comparable to VaEA in terms of IGD. Besides, the proposed algorithm is more robust and can expand the front better.

Keywords: Many-objective · Multiobjective · Evolutionary algorithm · EMO · NSGA-III

1 Introduction

Real-world problems often involve multiple objectives. For example, shop managers need to minimize makespan and tardiness in production scheduling problems, and fleet managers need to minimize the number of vehicles and total distance in vehicle routing problems. These concerned objectives are usually conflict with each other, which means that improvement on one objective leads to deterioration of other objectives. We are not able to define a single optimal solution when there is no prior information on the trade-off relationship between objectives. Pareto dominance helps to deal with this situation. It defines that a solution x *dominates* a solution y if and only if all objectives of x are not worse than those of y and at least one objective of x is better than that of y . If a solution x^* is not dominated by any solution, we call it a *Pareto optimal* solution. The set of all Pareto optimal solutions is called the *Pareto set*, and the projection of Pareto set onto the objective space is called the *Pareto front*. The goal of solving a multiobjective optimization problem (MOP) by a Pareto approach is to find or approximate the Pareto set and/or the Pareto front.

With the population-based nature, evolutionary algorithms (EAs) have the potential to solve the MOP effectively and efficiently. In the last two decades, evolutionary multiobjective optimization (EMO) has become an active and fast-growing research topic in the field of evolutionary computation. Well-known dominance-based multiobjective evolutionary algorithms (MOEAs) such as NSGA-II [1] and SPEA2 [2] triggered a lot of following research studies and applications. Dominance-based

selection, however, encounters challenges when the number of objectives gets larger and larger. Many-objective optimization problems (MaOP) refer to the problems with more than three objectives. As the number of objectives increases, the dominance relationship between solutions gets weak, i.e., it becomes difficult to assess solution quality based on the dominance relationship. Many studies have reported the difficulty of solving MaOPs [3, 4]. Decomposition-based algorithms are another category of algorithms. They decompose the MOP into multiple single objective optimization problems by some scalarizing functions (e.g. weighted sum or weighted Tchebycheff) or decompose the solution space into multiple regions by a set of weight vectors or reference points. MOEA/D [5] is a representative of the decomposition-based algorithms. The idea of decomposition inspires many following many-objective algorithms, such as NSGA-III, MOEA/DD [6], θ -DEA [7], VaEA [8], and so on.

In this paper, we focused on NSGA-III and proposed some modifications. The rest of this paper is organized as follows. Section 2 will review recent many-objective evolutionary algorithms, especially those taking both ideas of dominance and decomposition. Section 3 will present the proposed algorithm and modifications. Experiments and results are given in Sect. 4. Section 5 gives conclusions and lists future work.

2 Literature Review

After its proposal in 2002, the dominance-based algorithm NSGA-II dominated the field of EMO for a long time. It has been cited by about 24,000 times according to the statistics of google scholar. Its next version, NSGA-III, was finally proposed twelve years later in 2014. NSGA-III keeps the $(n + n)$ environmental selection and non-dominated sorting in NSGA-II. It has an adaptive normalization method and constructs a hyperplane on the normalized objective space. A set of reference points are then produced on the hyperplane. Each individual in the population is associated with a reference point based on the perpendicular distance to the reference line passing the point. The number of associated individuals of a reference point measures the density of the corresponding region. By favoring individuals in low-density regions, NSGA-III can maintain a well-distributed population. As commented in [6], NSGA-III “employs a decomposition-based idea to maintain population diversity, while the convergence is still controlled by Pareto dominance.”

The cooperation of dominance and decomposition was clearly realized and mentioned in MOEA/DD. MOEA/DD generates a set of weight vectors and associates each individual x with a weight vector w based on the angle between w and the objective vector $F(x)$. Each weight vector defines a subregion, and its niche count refers to the number of individuals in the subregion. MOEA/DD takes the $(n + 1)$ selection mechanism. After the non-dominated sorting of the n population members and the new offspring, one individual is removed according to the non-domination level of the individual, niche count of the subregion, sum of PBI (penalty-based boundary intersection) values of individuals in the subregion, and the PBI value of individual. A feature of MOEA/DD’s selection is that an individual in an isolated subregion might be kept even though it is dominated by another individual in the population.

θ -DEA [7] is similar to NSGA-III, also doing normalization, association, and non-dominated sorting. The main difference is that non-dominated sorting in θ -DEA relies on the proposed θ -dominance relationship, which is based on the PBI value in the normalized objective space. The θ -dominance relationship is defined only between individuals associated with the same reference point. There is also a slight difference from NSGA-III in the way they find the extreme points to construct the hyperplane.

VaEA [8] also uses the ideas of both dominance and decomposition. This study points out that generation of reference points is a challenge. First, the number of reference points increases very quickly as the number of objectives gets high; second, uniformly-distributed reference points cannot guarantee uniformly-distributed solutions when the problem to be solved has an irregular Pareto front. In this regard, VaEA was proposed as an optimizer based on the search directions of the population itself (instead of predefined directions). Like NSGA-III, the main steps of VaEA include non-dominated sorting, normalization, association, and niching. Its normalization is based on the ideal and nadir points, simpler than NSGA-III since it does not need hyperplane construction. The major difference is in that association in VaEA is based on the angle between collected individuals and uncollected individuals. The maximum-vector-angle-first principle collects from uncollected individuals the one has the largest minimal angle from the collected individuals. The worse-elimination principle replaces one collected individual by an uncollected individual if the angle between them is small and the uncollected individual has better convergence.

I-DBEA [9] is a decomposition-based algorithm but also considers dominance in the environmental selection. Its framework is similar to that of MOEA/D. I-DBEA also does adaptive normalization like NSGA-III, but it uses a different method to find the extreme points to construct the hyperplane. It removes the neighborhood in MOEA/D. Individuals are compared based on the PBI function but comparing hierarchically d_2 distance first then d_1 distance. If a child solution is not dominated by the population, it is compared with individuals in the population in a random order and replaces the first one that is worse than the child.

MaOEA-CSS [10] aims to coordinate the mating selection and environmental selection strategies. Mating selection is based on 2-tournament: the individual with a smaller achievement scalarizing function (ASF) value and a larger minimal angle from the other is the winner. If there is no clear winner, one is randomly selected. The winner serves as a parent in a probability proportional to the rank of its ASF value in the population. MaOEA-CSS removes individuals iteratively. Each time a pair of individuals with the minimal angle between each other is selected. They are evaluated by the Euclidean distance to the ideal point and the angle to others excluding them. When the difference between the distances is large, the one with the larger distance is removed; otherwise, the one with a smaller angle is removed.

As the name indicates, MOEA/D-DU [11] is an extension of MOEA/D and is a pure decomposition-based algorithm. The authors of MOEA/D-DU noticed that in MOEA/D a child may replace an individual in the population even though the child is far away from the corresponding weight vector (i.e. search direction). Thus, they proposed a new environmental selection mechanism in which a child tests if it can replace the individuals in the population in non-decreasing order of the perpendicular

distance to their corresponding weight vectors. Experimental results showed that MOEA/D-DU performed better than I-DBEA.

3 Improved NSGA-III Algorithm (I-NSGA-III)

There are two random selection in the original NSGA-III. First, NSGA-III randomly selects parents for reproduction. Second, in the niching procedure, NSGA-III randomly selects among the individuals in a cluster when the niche count is greater than zero. We propose modifications on these two random selection and do some investigation.

Table 1. Pseudo code of improved NSGA-III

01	Initialize (P, N)
02	for t = 1 to MaxGen
03	Q \leftarrow \emptyset
04	for i = 1 to N/2
05	{p1, p2} \leftarrow SelectParent (P)
06	{c1, c2} \leftarrow Reproduce (p1, p2)
07	Q \leftarrow Q \cup {c1, c2}
08	R \leftarrow P \cup Q, P \leftarrow \emptyset
09	Normalize (R)
10	NondominatedSort (P, Fl, R)
11	Associate (P \cup Fl)
12	Niching (P, Fl)

3.1 Initialization

Each individual X_i ($i = 1, 2, \dots, N$) is a real vector, where N is the population size. The initial population is generated randomly. Technically speaking, decision variables are set by random values in the problem-specified interval.

3.2 Mating Selection and Reproduction

Each individual in the population is associated with a reference point. (The association procedure will be described later in Sect. 3.5. The niche count of a reference point refers to the number of individuals associated with the point. The proposed mating selection mechanism does 2-tournament in two phases – selection of reference points and then selection of individuals. In the first phase the reference point with a smaller niche count is favored; in the second phase, among the individuals associated with the selected reference point, the one with a smaller Euclidean distance to the ideal point is favored. The SBX crossover and polynomial mutation are used in the reproduction step. Values of their parameters will be given in the Sect. 4.2.

3.3 Normalization

In our algorithm, we adopt the simpler normalization procedure in VaEA. The ideal point Z^{\min} and the nadir point Z^{\max} are determined by the minimal and maximal objective values in the current population. The j^{th} normalized objective is calculated by

$$f'_j(X_i) = (f_j(X_i) - Z_j^{\min}) / (Z_j^{\max} - Z_j^{\min}), j = 1, 2, \dots, M, \quad (1)$$

where M is the number of objectives.

3.4 Nondominated Sort

By applying the non-dominated sort procedure to the union of current population and the newly produced offspring (R in Table 1), these $2 \cdot N$ individuals are classified into one or several fronts. Putting these individuals front by front into the next population P until adding the last front Fl exceeds the population size. The union of P and Fl goes to the next step – association.

3.5 Association

We use Das and Dennis' approach [12] to generate a set of reference points in the normalized objective space. The line passing the ideal point and a reference point is called the reference line of the reference point. We calculate for each individual the perpendicular distance from each reference line, and then associate individuals in P and Fl with the corresponding reference point of the reference line having the shortest distance.

3.6 Niching

The niching procedure of NSGA-III adds individuals in Fl iteratively into the next population P to meet the population size. Let I_j denote the set of individuals in Fl associated with a reference point j . In each iteration, firstly the reference point j^* with the smallest niche count is selected. When its niche count is zero, from I_{j^*} the individual with the shortest perpendicular distance to the reference line is added. When its niche count is greater than zero, a random individual from I_{j^*} is added.

Ishibuchi et al. [13] pointed out that performance of decomposition-based many-objective optimization algorithms including NSGA-III strongly depends on the shape of Pareto front. Ishibuchi et al. modified DTLZ functions to the negative version and changed the shapes of Pareto front. Experimental results showed that performance of NSGA-III deteriorates on these negative DTLZ functions. Motivated by the finding in [13], we hybridized the idea of VaEA into NSGA-III. When the niche count of the selected reference point is greater than zero, we add the individual in Fl that has the largest minimal angle from the individuals in P . The angle between two individuals X_i and X_j is calculated by (2). The part of reference point-based selection of NSGA-III selects individuals in a systematic way at the beginning; then, the angle-based selection in VaEA considers the diversity and may select more suitable individuals than the part of random selection of NSGA-III does.

$$\text{angel}(X_i, X_j) = \arccos\left|\left(\frac{F'(X_i) \bullet F'(X_j)}{\text{norm}(X_i) \cdot \text{norm}(X_j)}\right)\right| \quad (2)$$

$$F'(X_i) \bullet F'(X_j) = \sum_{k=1}^M f'_k(X_i) \cdot f'_k(X_j) \quad (3)$$

$$\text{norm}(X_i) = \sqrt{\sum_{k=1}^M f'_k(X_i)^2} \quad (4)$$

4 Experiments and Results

4.1 Benchmark Functions and Algorithms

We compare performance of the proposed improved NSGA-III (called I-NSGA-III hereafter) with NSGA-III and VaEA on the negative version of DTLZ1-4 [13, 14] with the number of objectives $M \in \{3, 5, 8, 10\}$. Due to the limit of space, results of solving the 10-objective functions are omitted in Table 2. The negative version of DTLZ functions multiplies the original objective values by -1. This changes the shape and the size of the Pareto front. For the detailed definitions, please refer to [13].

4.2 Parameter Setting and Performance Measures

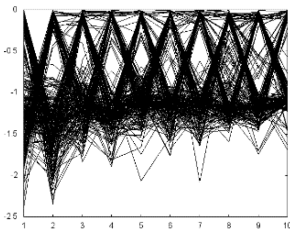
Parameter setting in the experiments is the same as that in [4, 8]. Crossover rate is 1.0, and mutation rate is $1/D$, where D is the number of decision variables. The parameter η_c in SBX is 30 and the parameter η_m in polynomial mutation is 20. Population size and generation number depend on the DTLZ function and the number of objectives. They are listed together with the performance results in Table 2. Performance is measured by inverted generational distance (IGD), which calculates the average Euclidean distance from the reference Pareto front to the set of solutions obtained by the tested algorithm. We generated the reference front by sampling 10,000 random points on the hyperplane $\sum x_i = -551.16$ for DTLZ1^{-1} , $\sum x_i^2 = (-3.5)^2$ for DTLZ2^{-1} and DTLZ4^{-1} , and $\sum x_i^2 = (-2203.7)^2$ for DTLZ3^{-1} , respectively.

4.3 Comparison Results

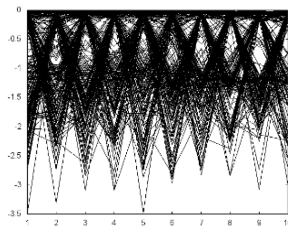
Each algorithm solved each of the 16 problems for 20 times. The mean and standard deviation of IGD values are listed in Table 2. We can see that the IGD performance of NSGA-III is improved by the proposed selection mechanism for all 16 problems. Although VaEA has the smallest mean IGD for 13 out of 16 problems, the difference between VaEA and I-NSGA-III is not large in many problems. I-NSGA-III has smaller standard deviation of IGD values than VaEA for all 16 problems, which reveals that the proposed idea improves the robustness. Besides, when we observed the obtained solutions of the three algorithms, we found that NSGA-III and VaEA are not able to expand the obtained front when solving DTLZ2^{-1} , DTLZ3^{-1} , and DTLZ4^{-1} . Figure 1

Table 2. Mean and standard deviation of IGD values over 20 runs

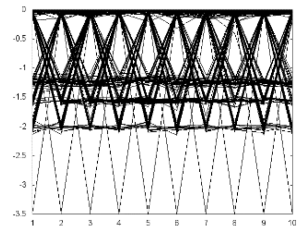
Problem	M	N	Gen. Num.	NSGA-III	VaEA	I-NSGA-III
DTLZ1 ⁻¹	3	92	400	3.180E + 1 (9.757E-1)	2.427E + 1 (5.343E-1)	2.369E + 1 (1.475E-1)
	5	210	600	7.136E + 1 (3.448E + 0)	5.391E + 1 (7.776E-1)	5.197E + 1 (4.009E-1)
	8	156	750	2.230E + 2 (1.126E + 1)	9.475E + 1 (1.125E + 0)	9.997E + 1 (6.816E-1)
DTLZ2 ⁻¹	3	92	250	2.353E-1 (4.339E-3)	2.243E-1 (4.701E-3)	2.291E-1 (3.226E-3)
	5	210	350	7.145E-1 (1.483E-2)	6.058E-1 (1.252E-2)	6.141E-1 (4.088E-3)
	8	156	500	1.699E + 0 (4.069E-2)	1.240E + 0 (1.356E-2)	1.257E + 0 (6.799E-3)
DTLZ3 ⁻¹	3	92	1000	1.519E + 2 (2.747E + 0)	1.433E + 2 (4.093E + 0)	1.508E + 2 (1.371E + 0)
	5	210	1000	4.549E + 2 (1.129E + 1)	3.939E + 2 (9.449E + 0)	3.899E + 2 (3.057E + 0)
	8	156	1000	1.053E + 3 (1.736E + 1)	7.851E + 2 (9.925E + 0)	7.948E + 2 (3.149E + 0)
DTLZ4 ⁻¹	3	92	600	2.394E-1 (5.443E-3)	2.291E-1 (6.57E-3)	2.393E-1 (1.188E-3)
	5	210	1000	7.382E-1 (1.611E-2)	6.141E-1 (8.005E-3)	6.293E-1 (3.692E-3)
	8	156	1250	1.375E + 0 (4.11E-2)	1.257E + 0 (1.535E-2)	1.279E + 0 (5.566E-3)



(a) NSGA-III



(b) VaEA



(c) I-NSGA-III

Fig. 1. Value path plot of obtained solutions in one run for DTLZ2⁻¹ with $M = 10$

illustrates the solutions obtained by each algorithm in a typical run of solving DTLZ2⁻¹ with $M = 10$. NSGA-III cannot reach the minimal value -3.5 for any objective, and VaEA can only reach -3.5 for some of 10 objectives. I-NSGA-III can reach -3.5 for all 10 objectives.

5 Conclusions

In this paper we focused on two random selection in NSGA-III and proposed modifications. In mating selection, we considered diversity by the niche count and convergence by the Euclidean distance to the ideal point. In environmental selection, we combined the reference point-based selection in NSGA-III and the vector angle-based selection in VaEA. Experimental results showed that the proposed I-NSGA-III outperformed the original NSGA-III on the negative version of DTLZ functions in terms of IGD. The IGD performance of I-NSGA-III is competitive when comparing with VaEA. Besides, I-NSGA-III can expand the front better than NSGA-III and VaEA. In our future work, we will study the detailed reason why the hybrid selection is able to expand the front better. We actually did some experiments to examine the effect of the proposed mating selection and found that the effect is not obvious. We will keep figuring out the reason. The idea of neighborhood in MOEA/D and some better reproduction operators will also be future research topics.

Acknowledgement. This research was supported by the Ministry of Science and Technology of Taiwan (R.O.C.) under Grant No. 105-2221-E-003-021.

References

1. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans. Evol. Comput.* **6**(2), 181–192 (2002). [NSGA-II]
2. Zitzler, E., Laumanns, M., Thiele, L.: SPEA2: Improving the strength Pareto evolutionary algorithm. Technical Report 103, Department of Electrical Engineering, Swiss Federal Institute of Technology (ETH), Zürich, Switzerland (2001). [SPEA2]
3. Ishibuchi, H., Tsukamoto, N., Nojima, Y.: Evolutionary many-objective optimization: a short review. In: *Proceedings of the IEEE Conference on Evolutionary Computation*, pp. 2419–2426 (2008)
4. Deb, K., Jain, H.: An evolutionary many-objective optimization algorithm using reference-point based non-dominated sorting approach, part I: solving problems with box constraints. *IEEE Trans. Evol. Comput.* **18**(4), 577–601 (2014). [NSGA-III]
5. Zhang, Q., Li, H.: MOEA/D: a multiobjective evolutionary algorithm based on decomposition. *IEEE Trans. Evol. Comput.* **11**(6), 712–731 (2007). [MOEA/D]
6. Li, K., Deb, K., Zhang, Q., Kwong, S.: An evolutionary many-objective optimization algorithm based on dominance and decomposition. *IEEE Trans. Evol. Comput.* **19**(5), 694–716 (2015). [MOEA/DD]
7. Yuan, Y., Xu, H., Wang, B., Yao, X.: A new dominance relation-based evolutionary algorithm for many-objective optimization. *IEEE Trans. Evol. Comput.* **20**(1), 16–37 (2016). [Theta-DEA]
8. Xiang, Y., Zhou, Y., Li, M., Chen, Z.: A vector angle-based evolutionary algorithm for unconstrained many-objective optimization. *IEEE Trans. Evol. Comput.* **21**(1), 131–152 (2017). [VaEA]
9. Asafuddoula, M., Ray, T., Sarker, R.: A decomposition-based evolutionary algorithm for many objective optimization. *IEEE Trans. Evol. Comput.* **19**(3), 445–460 (2015). [I-DBEA]
10. He, Z., Yen, G.: Many-objective evolutionary algorithms based on coordinated selection strategy. *IEEE Trans. Evol. Comput.* **21**(2), 220–233 (2017). [MaOEA-CSS]

11. Yuan, Y., Xu, H., Wang, B., Zhang, B., Yao, X.: Balancing convergence and diversity in decomposition-based many-objective optimizers. *IEEE Trans. Evol. Comput.* **20**(2), 180–198 (2016). [MOEA/D-DU]
12. Das, I., Dennis, J.: Normal-boundary intersection: A new method for generating the Pareto surface in nonlinear multicriteria optimization problems. *SIAM J. Optim.* **8**(3), 631–647 (1998)
13. Ishibuchi, H., Setoguchi, Y., Masuda, H., Nojima, Y.: Performance of decomposition-based many-objective algorithms strongly depends on Pareto front shapes. *IEEE Trans. Evol. Comput.* **21**(2), 169–190 (2017)
14. Deb, K., Thiele, L., Laumanns, M., Zitzler, E.: Scalable test problems for evolutionary multiobjective optimization. In: *International Conference on Evolutionary Multi-Criterion Optimization*, pp. 105–145. Springer (2005). [DTLZ]

Freeway Travel Time Prediction by Using the GA-Based Hammerstein Recurrent Neural Network

Ru-Kam Lee¹, Yi-Che Yang¹, Jun-Hong Chen¹,
and Yi-Chung Chen^{2(✉)}

¹ Department of Information Engineering and Computer Science,
Feng Chia University, Taichung, Taiwan, R.O.C.
kiam04@hotmail.com, a3104123@gmail.com,
ken83715@gmail.com

² Department of Industrial Engineering and Management,
National Yunlin University of Science and Technology, Yunlin, Taiwan, R.O.C.
mitsukoshi901@gmail.com

Abstract. Freeway travel time prediction has become a focus of research in recent years. However, we must understand that most conventional methods are very instinctive. They rely on the small amount of real-time data from the day of travel to look for historical data with similar characteristics and then use the similar data to make predictions. This approach is only applicable for a single day and cannot be used to predict the travel time on a day in the future (such as looking up the travel time for the coming Sunday on a Monday). This study therefore developed a Hammerstein recurrent neural network based on genetic algorithms that learns the freeway travel time for different dates. The trained model can then be used to predict freeway travel time for a future date. The experiment results demonstrated the validity of the proposed approach.

Keywords: Freeway travel time prediction · Recurrent neural network · Genetic algorithm

1 Introduction

Travel time prediction enables road users and traffic management to accurately predict travel time. This is needed because the travel time of the same route may vary greatly with what time of the day it is, whether said day is a holiday, and how the weather is. Conventional predictions use the real-time data of the travel day to look for days with similar data in the historical records and then using the data of those days to make predictions. However, this approach is only applicable for a single day and cannot be used to predict the travel time on a day in the future (such as the travel time during a long weekend), so a new method is needed to resolve this issue.

After reading relevant studies on travel time prediction, we found that existing methods rely heavily on historical data, so when the conditions of the historical data differ from those of the travel route in question, the prediction accuracy will be greatly decreased. For instance, suspended tolls and high-occupancy vehicle restrictions are often implemented in Taiwan during special holidays. Suppose that the government does

not implement these measures one year; past conditions will not be the same, and so historical data will not be of use. Furthermore, previous prediction methods based on historical traffic flow generally divided historical data into groups first [11] before comparing them with the current traffic flow. This approach generally fails to make accurate predictions when there are unpredictable traffic issues, such as automobile accidents. This is another issue that warrants the development of new prediction methods.

To overcome the issues of insufficient data for special circumstances and poor response to unexpected situations, we incorporated the genetic algorithm-based Hammerstein recurrent neural network (RNN) [10]. In the event of insufficient historical traffic flow data, the k -nearest neighbor algorithm generally needs several learning cycles before making effective predictions. In contrast, an RNN only needs one learning cycle and thus offers significantly greater efficiency. In existing travel time prediction methods, sudden incidents often make real-time data incomparable with historical data, which then results in inaccurate predictions. An RNN can use the results of the time before sudden incidents to effectively predict the changes in traffic flow after the incidents, thereby solving this issue. Furthermore, the goal of prediction is to obtain the optimal solution of the model without considering the training time of the model. Thus, we adopted a genetic algorithm rather than the conventional back propagation approach to train the target RNN. The simulation results in this study demonstrate the validity of the proposed approach.

2 Related Works

2.1 Works Regarding Freeway Travel Time Prediction

The majority of recent studies regarding freeway travel time prediction employed the k -NN method, which we introduce below. First developed by Benedetti [2], Stone [8], and Tukey [9] based on the concept of nearest neighbors, the k -NN method searches historical data for data with characteristics similar to those of the real-time data. Clark [3] employed this approach to conduct cross-analysis of traffic flow, occupancy rates, and velocity of vehicles, whereas Tsai [11] modified Clark's method, set different detector parameters, and added the dimension of time to develop an even more accurate model:

$$tss = \sum_{i=1}^L \sum_{j=1}^T \left[w_q \left(q_{ij}^r - q_{ij}^h \right)^2 + w_v \left(v_{ij}^r - v_{ij}^h \right)^2 \right], \quad (1)$$

where tss is the sum of the squares; L denotes the total number of detectors; T is the time length; w indicates a weight coefficient; q denotes traffic flow; v is velocity; r and h indicate real-time data and historical data, respectively; i denotes the time point, and j is the detection value.

2.2 Works Regarding Recurrent Neural Networks

The use of RNNs in place of time series modeling is a well-known category in the discipline of system identification [6, 10]. In this discipline, RNNs have been

demonstrated to be one of the most efficient methods to process complex and dynamic problems. However, no RNN algorithm or model has been truly recognized [7]. As numerous RNN structures exist, researchers generally find the most suitable RNN structure for a particular problem by trial-and-error, which is very time-consuming. Thus, researchers have begun developing automated RNNs in recent years to develop algorithms. Such algorithms can automatically complete the process of system identification and generally include effective parameter initialization methods and learning algorithms that can operate steadily with online parameters. In recent years, RNNs with block-oriented (BO) models have been considered the most suitable models to solve dynamic nonlinear problems. For instance, the Hammerstein, Wiener, or Hammerstein-Wiener models all comprise linear dynamic subsystems and nonlinear static subsystems and are widely applied in system identification. Westwick and Kearney [4] used the Hammerstein model to identify stretch reflex dynamic systems. Kalafatis et al. [1] successfully applied the Wiener model to PH processes.

3 Algorithm

This section introduces the proposed algorithm, including (1) the structural design of the eTag traffic data and (2) the RNN.

3.1 Structural Design of eTag Traffic Data

Format of raw data: Table 1 shows the raw data obtained from the traffic database of the Taiwan Area National Freeway Bureau of the Ministry of Transportation and Communications, which include time and date, origin, destination, type of vehicle, travel time, and traffic flow. From this table, we can see that at 12:55 pm on February 11, 2015, 46 vehicles of type 31 (Light vehicles) took 45 min to travel the route from detector 01F0017N to detector 01F0005N.

Table 1. Raw traffic data

Date and time	Origin	Destination	Vehicle type	Travel time (min)	Traffic flow (vehicles)
2017/2/11 12:55	01F0017N	01F0005N	31	45	46
2017/2/11 12:55	01F0017N	01F0005N	32	47	6
2017/2/11 12:55	01F0017N	01F0005N	41	48	2

Format of data for prediction calculation: Table 2 shows the data format converted from the raw data for the prediction calculations in this study, the fields of which include number, origin, destination, total traffic flow, average travel time, date, and time. The main differences were the added numbers to each tuple of data, the separation

of date and time to facilitate grouping later on, and the calculation of total traffic flow and average travel time for further analysis.

Table 2. Traffic data for prediction calculation

No.	Origin	Destination	Total traffic flow (vehicles)	Average travel time (min)	Date	Time
0001	01F0017N	01F0005N	54	46	2017/2/11	12:55
0002	03F0116N	01F0099N	9	36	2017/2/11	23:30
0003	01F0061S	01F0099S	62	152	2017/2/12	18:30

3.2 Hammerstein Recurrent Neural Network

In this next section, we explain the Hammerstein RNN developed in this study in detail, including the model design and the derivation of the model algorithm.

Model Design

In model design, a Hammerstein model [5] refers to a BO model with a static nonlinear subsystem in front and a dynamic linear subsystem in back. Based on this subsystem arrangement, we can design an RNN with the characteristics of a Hammerstein model, as shown in Fig. 1. This model has a four-layer framework, including the input layer, hidden layer, dynamic layer, and output layer. As the name suggests, the input layer is used to receive input data and then transfer the data received to the other layers in the neural network. The neuron in the middle does nothing. The hidden layer of the neural network is responsible for constructing the status nonlinear subsystem. The neurons of this layer have no recurrent items, which means that they are static. The nonlinear function used by the neurons in this layer is responsible for processing the nonlinear part of the static nonlinear subsystem. Next, the dynamic layer and the output layer are in charge of constructing the dynamic linear subsystem. The dynamic layer uses the recurrent items to process the dynamic part, and the output layer processes the linear part. Below, we use equations to introduce the structural details of our model. For the sake of convenience, we considered the most widely used tangent sigmoid function as the target nonlinear function of the neural network. In the neurons of layer j , we $u_i^{(j)}(k)$, $f_i^{(j)}(k)$, and $o_i^{(j)}(k)$ to represent the input of neuron i at time k , the input of the function, and the output.

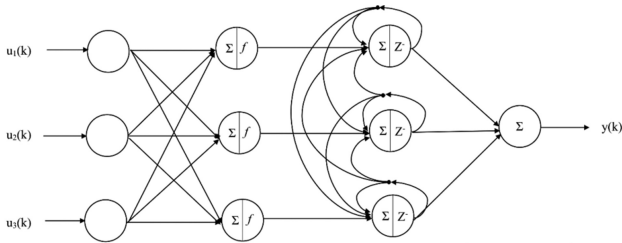


Fig. 1. The structure of Hammerstein recurrent neural network

First is the input layer. As the input layer does not do anything, its output is directly designated as the input:

$$o_i^{(1)}(k) = f_i^{(1)}(k) = u_i^{(1)}(k). \quad (2)$$

Next is the hidden layer, the input of which undergoes the calculations of the tangent sigmoid function. Thus, the equation of this layer can be expressed as

$$f_i^{(2)}(k) = \sum_{j=1}^p w_{ij} o_j^{(1)}(k) + d_i(k), \quad (3)$$

$$o_i^{(2)}(k) = \frac{\exp(f_i^{(2)}(k)) - \exp(-f_i^{(2)}(k))}{\exp(f_i^{(2)}(k)) + \exp(-f_i^{(2)}(k))}, \quad (4)$$

where w denotes the weight value between the input layer and the hidden layer; d is a bias value, and $\exp(\bullet)$ represents an exponential function.

Next in the dynamic layer, the neurons use their previous output and the current output of the hidden layer to calculate their current output. The equation can be written as

$$f_i^{(3)}(k+1) = \sum_{j=1}^q a_{ij} x_j(k) + \sum_{m=1}^p b_{im} o_m^{(2)}(k), \quad (5)$$

$$o_i^{(3)}(k) = x_i(k) = f_i^{(3)}(k). \quad (6)$$

Finally, the output layer integrates the data values from the dynamic layer and sends out the final results. Its equation is thus:

$$o_i^{(4)}(k) = f_i^{(4)}(k) = \sum_{j=1}^r c_{ij} x_j(k). \quad (7)$$

Use of Genetic Algorithm to Train Proposed Recurrent Neural Network

We used a genetic algorithm to train the proposed recurrent neural network. Of particular note, we used a genetic algorithm rather than the conventional back propagation algorithm for training because genetic algorithms can help us find the global optimum of the proposed neural network, whereas back propagation cannot. During our training process, we changed each weight value, including \mathbf{w} , \mathbf{d} , \mathbf{a} , \mathbf{b} , and \mathbf{c} , into 8-bit 01 series. The fitness function of the evolution process can be written as

$$Error(\mathbf{w}, i) = 1/2(y_d(i) - y(i))^2 = 1/2error(i)^2, \quad (8)$$

where $error(i)$ represents the error between the ideal output and the network output. With the settings of these two items and the genetic algorithm, we can complete the training of the target RNN.

4 Simulation Results

We examined three routes in this study: Yuanshan-Donghu, Fengyuan-Houli, and Tainan-Madou, chosen for their different traffic flows distributions. The Yuanshan-Donghu route has evenly distributed traffic flows throughout the day, the Fengyuan-Houli route is generally congested in the mornings and afternoons, and the traffic flow in the Tainan-Madou route is greatest around noon. These different traffic flows distributions

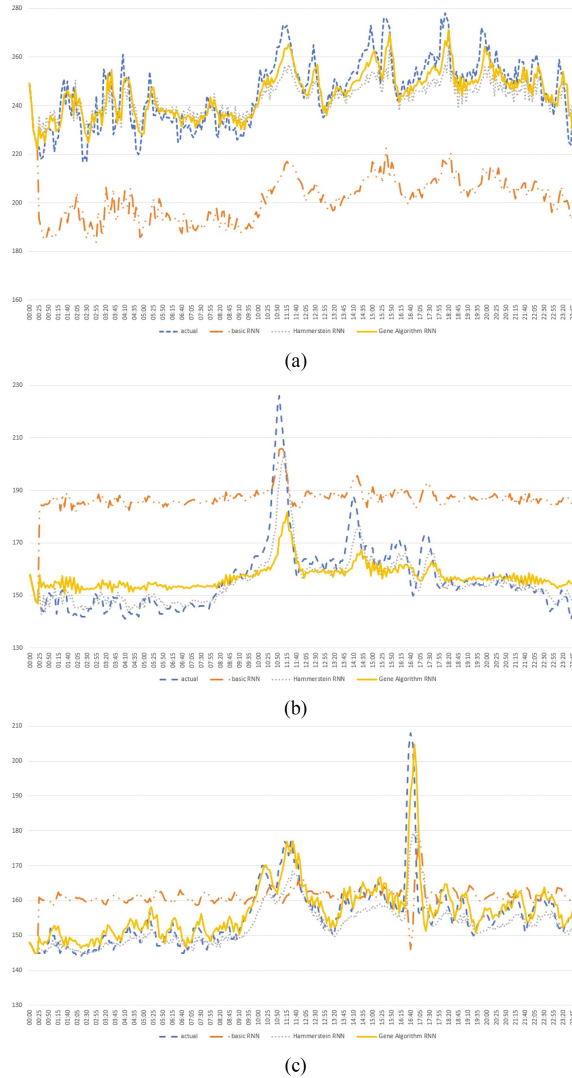


Fig. 2. Comparison of proposed and conventional methods; (a) Yuanshan-Donghu, (b) Fengyuan-Houli, (c) Tainan-Madou

Table 3. Comparison of proposed and conventional methods (MSE)

(MSE)	Basic RNN	Hammerstein RNN + BP	Hammerstein RNN + GA
Yuanshan-Donghu	45.6154	0.0087	0.0066
Fengyuan-Houli	32.2171	0.0087	0.0145
Tainan-Madou	11.0755	0.0178	0.0138

allowed us to verify the performance of the proposed model. We designed a basic RNN and a Hammerstein-based RNN+ back propagation training algorithm to simulate the three different traffic flow patterns and compared the results with those of a Hammerstein-based RNN+ GA training algorithm. The raw data and a comparison of the results of the two methods are presented in Fig. 2. To help the readers compare the two methods, we list the errors of the three methods in Table 3. As can be seen, the Hammerstein model produces significantly better prediction results than the conventional approach.

5 Conclusions

Existing methods used to predict freeway travel time are generally complex and inaccurate and rely heavily on historical data. Thus study therefore proposed a recurrent neural network based on the Hammerstein model using system identification methods for freeway travel time prediction. The proposed approach is significantly more accurate than conventional methods. However, as there are many types of BO models, we will experiment with other models in the future and make further improvements on prediction performance based on the approach proposed in this study.

Acknowledgement. This work was supported in part by the Ministry of Science and Technology of Taiwan, R.O.C., under Contracts MOST 105-2119-M-035-002 and MOST 105-2634-E-035-001. Also, We are grateful to the National Center for High-performance Computing in Taiwan for computer time and facilities.

References

1. Kalafatis, A.D., Wang, L., Cluett, W.R.: Linearizing feedforward-feedback control of PH processes based on the Wiener model. *J. Process Control* **15**, 103–112 (2005)
2. Benedetti, J.: On the nonparametric estimation of regression functions. *J. Roy. Stat. Soc. B* **39**, 248–253 (1977)
3. Clark, S.: Traffic prediction using multivariate nonparametric regression. *J. Transp. Eng.* **129** (2), 161–168 (2003)
4. Westwick, D.T., Kearney, R.E.: Separable least squares identification of nonlinear Hammerstein models: application to stretch reflex dynamics. *Ann. Biomed. Eng.* **29**, 707–718 (2001)
5. Wang, J.S., Hsu, Y.L.: An MDL-based Hammerstein recurrent neural network for control applications. *Neurocomputing* **74**, 315–327 (2010)

6. Wang, J., Zhang, Y., Gao, Y., Xing, C.: PLSM: a highly efficient LSM-tree index supporting real-time big data analysis. In: Proceedings of International Conference on Computer Software and Applications, pp. 240–245 (2013)
7. Gori, M., Mozer, M., Tsoi, A.C., Watrous, R.L.: Presenting the special issue on recurrent neural networks for sequence processing. *Neurocomputing* **15**, 181–182 (1997)
8. Stone, C.: Consistent nonparametric regression. *Ann. Stat.* **5**, 595–645 (1977)
9. Tukey, J.: *Exploratory Data Analysis*. Addison Wesley, Reading (1977)
10. Chow, T.W.S., Fang, Y.: A recurrent neural-network-based real-time learning control strategy applying to nonlinear systems with unknown dynamics. *IEEE Trans. Ind. Electron.* **45**(1), 151–161 (1998)
11. Tsai, C.K.: Freeway travel time prediction by using the k-NN method and comparison of different data classification. Master thesis of Department of Transportation & Logistics Management, National Chiao Tung University (2009)

A PIP-Based Approach for Optimizing a Group Stock Portfolio by Grouping Genetic Algorithm

Chun-Hao Chen^(✉) and Chih-Hung Yu

Department of Computer Science and Information Engineering,
Tamkang University, Taipei, Taiwan
chchen@mail.tku.edu.tw, allendose800901@hotmail.com

Abstract. Recently, some approaches have been proposed for finding a group stock portfolio (GSP). However, stock price series of stocks which are useful information may not be considered in those approaches. Hence, this study takes stock price series into consideration and presents a perceptually important point (PIP)-based approach for obtaining a GSP. Since the PIP is used, the proposed approach can handle stock price series with different lengths, which means that a more useful GSP could be found and provided to investors. Each chromosome is encoded by grouping, stock, and stock portfolio parts. To measure the similarity of series in groups, the series distance is designed and used as a part of fitness function. At last, experiments were conducted on a real dataset to show the advantages of the proposed approach.

1 Introduction

Since financial markets have many financial instruments and derivatives, including stocks, futures, and options [6, 9, 10], users have many choices when creating a portfolio. The goal of portfolio selection is to minimize the value at risk (VaR) and maximize the return on investment (ROI). Since many factors affect the profit of the portfolio, a sophisticated approach for deriving a portfolio that considers various factors is needed. The well-known approach for acquiring a stock portfolio is the mean-variance (M-V) model [8]. Investors can use the M-V model to find a portfolio with the maximum ROI or minimum VaR. Based on the M-V model, many evolutionary algorithms have been proposed for deriving a portfolio [1, 4, 5, 7].

Although there are lots of stock portfolio optimization approaches have been proposed, they are designed for finding a stock portfolio. However, only a stock portfolio is not enough because investors may not buy the suggested stocks due to some reasons. Hence, in the previous approach, an approach that can divide n stocks into K groups for deriving a group stock portfolio (GSP) is proposed by the grouping genetic algorithm (GGA) [3]. A GSP consists of a set of stock groups, and each stock group has a set of stocks. Thus, stocks in the same group indicate that they are similar.

Time series data is commonly seen in real application. Each data point in a time series means that a value at a certain time. Stock price series is a kind of times series, and it contains many useful information. However, stock price series do not be

considered in the previous approach [3]. Hence, the motivation of this paper is to take stock price series into consideration to improve the similarity of stock price series in groups for a GSP.

Since the length of stock price series may different, to achieve the goal, the perceptually important point (PIP) presented in [2] is utilized, and a PIP-based approach is proposed for optimizing a GSP by GGA in this paper. The proposed approach first encodes a GSP into a chromosome by grouping part, stock part and stock portfolio part according to [3]. In the previous approach, a chromosome is evaluated by the portfolio satisfaction and the group balance. The portfolio satisfaction is used to evaluate the profit and satisfaction of user's requests, and group balance is used to make groups in a chromosome have as similar number of stocks as possible. In this paper, to avoid high risk stocks, based on cash dividends, the stability factor is added to original portfolio satisfaction, namely the modified portfolio satisfaction in the proposed approach. The series distance is also developed to evaluate the similarity of stock price series in groups. Thus, a chromosome is evaluated by the modified portfolio satisfaction, the group balance and the series distance. Genetic operators are then executed to generate new offspring. The process is repeated until reaching stop conditions. At last, experiments on a real data were conducted to show the merits of the proposed approach.

2 Perceptually Important Point

The perceptually important point (PIP) is an important method for time series compression, and was proposed by Chung et al. [2]. Let a time series $T = \{t_1, t_2, \dots, t_n\}$. Below, an example is given to illustrate how the PIP works if the number of desired points is five. Firstly, it selects two points (t_1 and t_n) that are the first and last points of the series. The third point is the one in T with the maximum distance from the point to the line connected from t_1 to t_n , where the maximum distance is the vertical distance between a test point and the line connecting the two adjacent important points. The same process will be executed until desired five points are extracted. The process to find the five points is shown in Figure example is shown in Fig. 1.

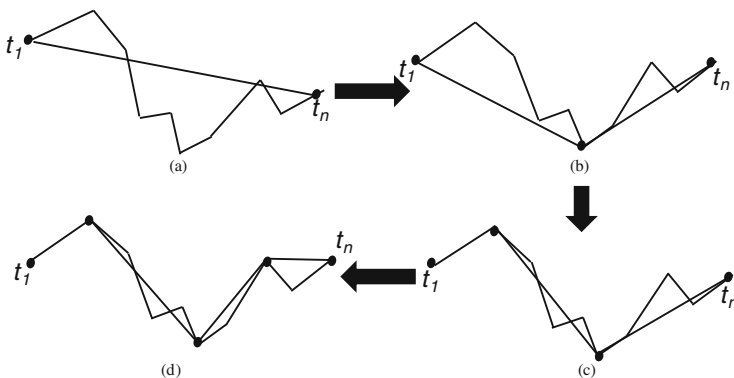


Fig. 1. An example for extracting important points by PIP.

3 The Proposed Approach

In this section, the PIP-based algorithm to mine group stock portfolio by grouping genetic algorithm is described. The details of the proposed algorithm are stated as follows:

- Input:** A set of stocks $S = \{s_i \mid 1 \leq i \leq n\}$ with their stock price series $T_i = \{d_{i1}, d_{i2}, \dots, d_{ih} \mid 1 \leq i \leq n\}$, a predefined maximum number of purchased stocks $numCom$, a predefined maximum investment capital $maxInves$, a predefined maximum number of purchased units of a stock $maxUnit$, cash dividends of stocks $Y = \{y_i \mid 1 \leq i \leq n\}$, a number of groups K , desired number of important points $desImpPoint$ parameters α , β and γ , a population size $pSize$, a crossover rate p_c , a mutation rate p_m , a inversion rate p_I , and a number of generations $generation$.
- Output:** The group stock portfolio GSP
- Step 1:** Generate initial population with $pSize$ using the given cash dividends Y .
- Step 2:** Transform each stock price series $T = \{d_{i1}, d_{i2}, \dots, d_{ih}\}$ to the desired number of points $desImpPoint$ by PIP and represent as $T' = \{t_1, t_2, \dots, t_{desImpPoint}\}$.
- Step 3:** Calculate fitness value of each chromosome using the following sub-steps:
- Sub-step 3.1:** Calculate modified portfolio satisfaction of a chromosome C_q using following formula:

$$PS(C_q) = \sum_{p=1}^{NC} subPS(SP_p) / NC,$$

where NC is the number of stock portfolios can be generated from a chromosome C_q , and $subPS(SP_p)$ is portfolio satisfaction of the p -th stock portfolio SP_p . The $subPS(SP_p)$ is calculated using following formula

$$subPS(SP_p) = \frac{ROI(SP_p)}{suitability(SP_p) * SF(SP_p)},$$

where the $ROI(SP_p)$ is profit of the stock portfolio in SP_p , the $suitability(SP_p)$ consists of the investment capital penalty and portfolio penalty, and the $SF(SP_p)$ is designed based on the cash dividend of stocks because it means the company is stable when cash dividends of a company are stable.

- Sub-step 3.2:** Calculate group balance of each chromosome using the formula:

$$GB(C_q) = \sum_{i=1}^K -\frac{|G_i|}{N} \log \frac{|G_i|}{N},$$

where $|G_i|$ is the number of stocks in the i -th group. Thus, when a chromosome has large group balance value, it means that numbers of stocks in groups are similar.

Sub-step 3.3: Calculate series distance of each chromosome using formula:

$$SD(C_q) = \frac{\sum_{i=1}^K \sum_{T_m^{G_i} \neq T_n^{G_i}} [EuclideanDist(F_m^{G_i}, F_n^{G_i})] / |F_n^{G_i}|}{\sum_{i=1}^K \sum_{T_m^{G_i} \neq T_n^{G_i}} 1},$$

where $F_m^{G_i}$ and $F_n^{G_i}$ are two series transformed by the PIP from stock price series $T_m^{G_i}$ and $T_n^{G_i}$ in group G_i , and can be represented as $F_m^{G_i} = \{f_{m1}, f_{m2}, \dots, f_{mk}\}$ and $F_n^{G_i} = \{f_{n1}, f_{n2}, \dots, f_{nk}\}$. The $EuclideanDist(F_m^{G_i}, F_n^{G_i})$ is the Euclidean distance of $F_m^{G_i}$ and $F_n^{G_i}$.

Sub-step 3.4: Set fitness value of each chromosome C_q using formulas:

$$f_1(C_q) = \frac{PS(C_q) * GB(C_q)^\alpha}{SD(C_q)^\gamma}.$$

- Step 4: Execute reproduction operation on the population to form the next population.
- Step 5: Execute crossover operation on the population.
- Step 6: Execute mutation operation on the population.
- Step 7: Execute inversion operation on the population.
- Step 8: If the stop criterion is satisfied, go to the next step. Otherwise, go to Step 3.
- Step 9: Output the chromosome with the best fitness value.

4 Experimental Results

In this section, experiments were made to show the merits of the proposed approach. The experimental dataset contains data from 2012/01/01 to 2014/12/31 that were collected from the Taiwan Stock Exchange (TSE). The dataset has 31 stocks, and the attributes include the stock prices, cash dividends and the risk values of stocks.

The average ROI, maximum ROI and minimum ROI of the derived GSPs are calculated according to the given training or testing datasets. The ten-run average return results of the proposed approach and previous approach [3] using one-year (2013) as training period and one-year (2014) as testing period are shown in Table 1.

Table 1. The average returns of the proposed and previous approaches on two-year training and testing datasets.

Training period (2013)	Avg. ROI	Max. ROI	Min. ROI
The previous approach	0.6	0.843	0.304
The proposed approach	0.445	0.692	0.163
Testing period (2014)	Avg. ROI	Max. ROI	Min. ROI
The previous approach	-0.009	0.186	0.304
The proposed approach	0.143	0.281	-0.033

Table 1 shows that the average ROIs (returns) of the previous and proposed approaches are 0.6 and 0.445, which means that they can reach good returns in training period. However, in the testing period, the average ROI of the proposed approach is 0.143 which is better than the previous approach. The results indicate that the proposed approach is effective to derive a better GSP when the new fitness function is used.

5 Conclusion and Future Work

To provide functional stock portfolios, an approach was presented to find a GSP by GGA in the previous approach. The aim is to divide stocks into groups and stocks in the same group means they are similar. However, the stock price series in groups of a GSP derived by previous approach may not similar. Thus, this paper takes stock price series of stocks into consideration and proposes a more sophisticated approach for obtaining a GSP which can not only improve the similarity of groups but also increase its return. Experiments were conducted on a real dataset to verify effectiveness of the proposed approach. In the future, some directions can also be extended based on the proposed approach. For example, the proposed approach could be verified on large dataset, or the island-based GA could be utilized to speed up the optimization process.

Acknowledgments. This research was supported by the Ministry of Science and Technology of the Republic of China under grant MOST 106-2221-E-032-049-MY2.

References

1. Bevilacqua, V., Pacelli, V., Saladino, S.: A novel multi objective genetic algorithm for the portfolio optimization. In: *Advanced Intelligent Computing*, pp. 186–193 (2012)
2. Chung, F.L., Fu, T.C., Luk, R., Ng, V.: Flexible time series pattern matching based on perceptually important points. In: *International Joint Conference on Artificial Intelligence Workshop on Learning from Temporal and Spatial Data*, pp. 1–7 (2001)
3. Chen, C.H., Lin, C.B., Chen, C.C.: Mining group stock portfolio by using grouping genetic algorithms. In: *The IEEE Congress on Evolutionary Computation*, pp. 738–743 (2015)
4. Chang, T.J., Yang, S.C., Chang, K.J.: Portfolio optimization problems in different risk measures using genetic algorithm. *Expert Syst. Appl.* **36**, 10529–10537 (2009)

5. Hoklie, L.R.Z.: Resolving multi objective stock portfolio optimization problem using genetic algorithm. In: International Conference on Computer and Automation Engineering, pp. 40–44 (2010)
6. Kumar, R., Bhattacharya, S.: Cooperative search using agents for cardinality constrained portfolio selection problem. *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.* **42**, 1510–1518 (2012)
7. Lin, P.C.: Portfolio optimization and risk measurement based on non-dominated sorting genetic algorithm. *J. Indus. Manag. Optim.* **8**, 549–564 (2012)
8. Markowitz, H.: Portfolio selection. *J. Finance* **7**(1), 77–91 (1952)
9. Ucar, I., Ozbayoglu, A.M., Ucar, M.: Developing a two level options trading strategy based on option pair optimization of spread strategies with evolutionary algorithms. In: *IEEE Congress on Evolutionary Computation*, pp. 2526–2531 (2015)
10. Wu, M.E., Wang, C.H., Chung, W.H.: Using trading mechanisms to investigate large futures data and their implications to market trends. *Soft. Comput.* **21**(11), 2821–2834 (2017)

A Novel Genetic Algorithm for Resource Allocation Optimization in Device-to-Device Communications

Yung-Fa Huang¹(✉), Tan-Hsu Tan², and Bor-An Chen²

¹ Department of Information and Communication Engineering,
Chaoyang University of Technology, Taichung 41349, Taiwan, R.O.C.

yfahuang@mail.cyut.edu.tw

² Department of Electrical Engineering,
National Taipei University of Technology, Taipei, Taiwan, R.O.C.
thtan@ntut.edu.tw

Abstract. In this study, the resource blocks (RB) are allocated to user equipment (UE) according to the evolutionary algorithms for long term evolution (LTE) systems. Genetic algorithm (GA) is one of the evolutionary algorithms, based on Darwinian models of natural selection and evolution. Therefore, we propose a novel GA for RB allocation to enhance the throughput of UEs and improve the system capacity performance. Simulation results show that the proposed GA with 100 populations, in 200 generations can converge to sub-optimal solutions. Therefore, with comparing with the particle swarm optimization (PSO) algorithm the proposed GA can improve system capacity performance with 1.4 UEs.

Keywords: Device-to-device · LTE communication systems · Genetic algorithm · Resource allocation

1 Introduction

The Orthogonal frequency division multiple access (OFDMA) scheme in the fourth generation mobile communication system (4G) long term evolution (LTE) technology not only upgrades spectrum efficiency but also provides high resistance for frequency-selective fading channels [1]. The Device-to-Device (D2D) is a developing key technique for next generation (5G) mobile communication systems [2].

In this study, the resource blocks (RB) are allocated to equipments (UEs) according to the evolutionary algorithms for LTE systems. The advantages of D2D techniques are such as improving energy efficiency of transmission, energy saving [4] and coverage rate improvement [5]. With relay station for transmission, most UEs can reach the minimum requirements of transmission rate [6]. Therefore, the application of relay station for transmission in the serious interference environment, the performance of system outperforms the directly transmission systems.

In previous works [7], the particle swarm optimization (PSO) algorithm was developed for resources allocation to achieve quality of service (QoS) and to maximize system capacity. The GA is a stochastic search algorithm whose procedures are based on the Darwinian models of natural selection and evolution [8]. Given some arbitrary initial solutions, the GA will generate the better solution through a series of genetic operations including selection, crossover, and mutation. Furthermore, The GA searches the solution space in parallel, that is, a set of possible solutions are manipulated in the same generation, so multiple local optimum can be reached simultaneously and thereby the likelihood of finding the global optimum is increased.

In previous works [9], we proposed a novel redundancy-saving genetic algorithm (RSGA) based on the cost value of the fitness function to improve the complexity in term the BER requirement for DS-CDMA systems. However, it is not suitable for OFDMA systems. Therefore, in this study, we further proposed a novel GA for the RB allocation for D2D systems.

2 System Model

In this paper the LTE communication systems based resources allocation issues are investigated. In this hybrid system, D2D UEs and traditional Cellular user equipments (CeUEs) share all resources. In the D2D communications, totally three relay stations are deployed in the cellular system, where UEs are uniformly random distributed in relay stations. In this study, base station is located in the center, and three relay stations form a triangle around the base station. Each relay station signal coverage with radius of 100 m. The CeUEs are uniformly random distributed in each relay stations within range. Each D2D pair includes a receiver UE and a transmitter UE distributed with 80 m distance from relay station. Channel models include Raleigh Fading, Shadowing Fading and path loss [10].

In the D2D communications, each uplink data transfer includes two hops. In the first hop, the u_l -th UE transmits signal to the l th relay station by channel gain $h_{u_l,l}^{(n)}$. The l th relay station relays the transmission to base station using the n th resource block (RB). However, when the u_j -th D2D UE transmits signal to the l th relay station, it will occurs the interference link gain $g_{u_j,l}^{(n)}$ to the l th relay station.

In this study, the channel models include path loss and shadowing fading. Thus, the fading channel model from UE to relay station (UE-relay) can be expressed by

$$PL_{u_l,l}(\ell)_{[\text{dB}]} = 103.8 + 20.9 \log(\ell) + L_{su} + 10 \log(\zeta) \quad (1)$$

where 103.8 is antenna gain; L_{su} is shadowing fading with log-normal distributed random variable where the standard deviation $\sigma = 10$; $10 \log(\zeta)$ is Rayleigh fading effect, ℓ is the distance between UE to relay station. Similarly, the fading channel model from relay station to BS (relay-BS) can be expressed by [10]

$$PL_{l,eNB}(\ell)_{[\text{dB}]} = 100.7 + 23.5 \log(\ell) + L_{su} + 10 \log(\zeta) \quad (2)$$

It is assumed that Base Station (BS) has know well the Channel State Information (CSI) of all channels. Then, the unit Power Signal-to-Interference-Plus-Noise Ratio (Unit Power SINR) of the first hop can be expressed by

$$\gamma_{u_l,l,1}^{(n)} = \frac{P_{u_l,l}^{(n)} h_{u_l,l}^{(n)}}{\sum_{\substack{\forall u_j \in U_j, \\ j \neq l, j \in L}} x_{u_j} P_{u_j,l}^{(n)} g_{u_j,l}^{(n)} + \sigma^2} \quad (3)$$

where $h_{u_l,l}^{(n)}$ is the channel gain from the u_l -th UE to the l th relay with the n th RB. $x_{u_l} = 1$ or 0. Each UE can only use one RB, $x_{u_l} = 1$ indicates with one RB, however, $x_{u_l} = 0$ indicates without any RB. U_j is the set of D2D UEs in the j th relay area. $P_{u_l,l}^{(n)}$ and $P_{u_j,j}^{(n)}$ are transmission power of the u_l -th UE and the u_j -th CeUE, respectively. $g_{u_j,l}^{(n)}$ is the interference link gain from the u_j -th CeUE to the l th relay. $\sigma^2 = N_0 B_{RB}$. N_0 is power spectral density of the added white Gaussian noise (AWGN). B_{RB} is bandwidth of an RB. Similarly, the unit power SINR of the second hop can be expressed by

$$\gamma_{l,eNB,2}^{(n)} = \frac{P_{l,eNB}^{(n)} h_{l,eNB}^{(n)}}{\sum_{\substack{\forall u_j \in U_j, \\ j \neq l, j \in L}} x_{u_j} P_{j,eNB}^{(n)} g_{j,eNB}^{(n)} + \sigma^2} \quad (4)$$

where $h_{l,eNB}^{(n)}$ is the channel gain from the l th relay to the Base Station (eNB) with the n th RB. $P_{l,eNB}^{(n)}$ and $P_{j,eNB}^{(n)}$ are transmission power of the l -th relay and the j -th relay, respectively. $g_{j,eNB}^{(n)}$ is the interference link gain from the j th relay to the eNB with the n th RB.

In this study, there are two links in the second hop. Equation (4) expresses the SINR of the l th relay station to BS. Similarly, the SINR of the l th relay station to the receiver of D2D pair can be expressed by

$$\gamma_{l,u_l,2}^{(n)} = \frac{P_{l,u_l}^{(n)} h_{l,u_l}^{(n)}}{\sum_{\substack{\forall u_j \in U_j, \\ j \neq l, j \in L}} x_{u_j} P_{j,u_l}^{(n)} g_{j,u_l}^{(n)} + \sigma^2} \quad (5)$$

In Eqs. (3)–(5), the SINR for all the links will be obtained. Then the throughput (Kbps) of the first hop can be derived by

$$r_{u_l,1}^{(n)} = B_{RB} \log_2(1 + \gamma_{u_l,l,1}^{(n)}) \quad (6)$$

Similarly, the throughput of the second hop can be obtained by

$$r_{u_l,2}^{(n)} = B_{\text{RB}} \log_2(1 + \gamma_{l,u_l,2}^{(n)}) \quad (7)$$

Thus, the throughput for the n th RB can be obtained by

$$R_{u_l}^{(n)} = \frac{1}{2} \min \{ r_{u_l,1}^{(n)}, r_{u_l,2}^{(n)} \} \quad (8)$$

With Eq. (8), the total throughput of N RBs is $\sum_{l \in L} \sum_{u_l \in U_l} \sum_{n=1}^N x_{u_l}^{(n)} R_{u_l}^{(n)}$. All UEs are desired to obtain the RBs to reach the maximal throughput. Therefore, to avoid deteriorating the communication quality a threshold Q_{u_l} is set to meet the required throughput for most UEs. Some constraints are set to perform the optimization problem as

$$\text{s.t.} \begin{cases} \sum_{u_l \in U_l} x_{u_l}^{(n)} \leq 1, \forall n \in N & \text{(a)} \\ \sum_{n=1}^N x_{u_l}^{(n)} P_{u_l,l}^{(n)} \leq P_{u_l}^{\max}, \forall u_l \in U_l & \text{(b)} \\ \sum_{n=1}^N x_l^{(n)} P_{l,u_l}^{(n)} \leq P_l^{\max}, \forall u_l \in U_l & \text{(c)} \\ R_{u_l} \geq Q_{u_l}, \forall u_l \in U_l & \text{(d)} \end{cases} \quad (9)$$

where Eq. 9(a) set the constraints for that each UE use only one RB. Equations 9(b) and (c) are the constraints on the minimum power for UEs and relay stations, respectively. Equation 9(d) is the minimum throughput requirements of QoS for UEs.

Moreover, the interference of the first hop and the second hop for system is expressed by

$$I_{u_l,1}^{(n)} = \sum_{\substack{\forall u_j \in U_j, \\ j \neq l, j \in L}} x_{u_j} P_{u_j,l}^{(n)} g_{u_j,l}^{(n)} \quad (10)$$

and

$$I_{u_l,2}^{(n)} = \begin{cases} \sum_{\substack{\forall u_j \in U_j, \\ j \neq l, j \in L}} x_{u_j} P_{j,eNB}^{(n)} g_{j,eNB,u_l}^{(n)}, u_l \in \{C \cap U_l\} \\ \sum_{\substack{\forall u_j \in U_j, \\ j \neq l, j \in L}} x_{u_j} P_{j,u_j}^{(n)} g_{j,u_l}^{(n)}, u_l \in \{D \cap U_l\} \end{cases}, \quad (11)$$

respectively.

In the system model, it is assumed that the base station coverage area is a circle with the radius 175 m. The D2D pairs are deployed on a circle with radius 80 m around each relay station. The numbers of UEs in the relay station are the same. It is assumed that the CSI of the links are known to base station.

3 RB Allocation with GA

In genetic algorithms (GAs) [8], the main idea is to follow the fittest evolutionary laws of nature, by the procedures of selection, crossover and mutation to improve the fitness value of chromosomes. With GA, there are random search, and other ways to search for the optimal solution. Therefore, the GA is often used to apply on optimization issues. In this study, the GA is applied for resources allocation optimization, with the objective functions on maximal system capacity and throughput.

The procedures in GA are the followings: (1) data coding, (2) producing initial population, (3) calculation fitness values, (4) selection, (5) crossover, and (6) mutation. The procedures are proceeded iterated from (3) to (6), until meeting the terminated conditions. Then the solution are obtained as the optimal results.

In the parent group, it is in accordance with the fitness value of chromosomes, to determine whether it will be retained or eliminated. In the select operation in this study, the ranking method, ranks the fitness value of each chromosome. This method can avoid inbreeding [11].

The crossover of GA is by selecting two chromosomes from the mating pool, and swapping the genes into two new chromosomes. It is expected that crossover procedures can generate better offspring chromosomes. Higher crossover rate in GA will bring the higher evolutionary rate for the chromosomes. Figure 1 shows a single point crossover process where from two selected chromosomes, A and B, to crossover two genes in to the two new chromosomes, C and D.



Fig. 1. Single point crossover of two chromosomes.

The mutation can increase the ethnic diversity of GA operations. The aforementioned selection, crossover and other procedures in both groups search for better children, but its genetic characteristics must be associated with the parent. Because there are no new chromosomes joining the group in each generation, it makes that the searching area cannot be expanded. It will lead the evolution to converge earlier. However, through mutation, some new chromosomes will join the search space to avoid GA early convergence problems.

In GA the object function is defined by

$$F_{obj_c}() = \sum_{u_i=1}^K y_{u_i} \quad (12)$$

where $\sum_{u_i=1}^K y_{u_i}$ is the system capacity and y_{u_i} is defined by

$$y_{u_i} = \begin{cases} 1, & \sum_{n=1}^N x_{u_i}^{(n)} R_{u_i}^{(n)} \geq R_{th} \\ 0, & \text{otherwise} \end{cases} \quad (13)$$

4 Simulation Results

The simulation parameters based on GA are shown in Table 1.

Table 1. Simulation parameters.

System bandwidth (MHz)	2.5
Bandwidth of subcarrier (Hz)	1500
Number of RBs	13
Radius of the coverage area of relay station (meter)	100
Distance between base station and relay station (meter)	125
Number of CeUE	9, 12, 15
Number of D2D pairs	3, 6, 9
Minimum distance between base station and UEs (meter)	10
Power of relay station (P_l , dBm)	30
Power of UEs (P_{u_i} , dBm)	23
Minimum throughput requirements of CeUEs (R_{th_C} , Kbps)	128
Minimum throughput requirements of D2D pairs (R_{th_D} , Kbps)	256
Standard deviation of Shadowing fading between relay and BS (dB)	6
Standard deviation of Shadowing fading between UEs and relay station (dB)	10
Power Spectral density of AWGN (dBm/Hz)	-174
Maximal generations (G)	200
Number of chromosomes (M)	10–500
Crossover rate (R_c)	0.9
Mutation rate (R_m)	0.07
Number of user equipments (K) (Including CeUE and D2D UEs)	12, 18, 24
Number of relay station	3
ID of RBs	1–13

Figure 2 shows that the system capacity for different population size in GA with $K = 18$. The objective function is $f_{obj_C}()$ in Eq. (12). When $M = 10$, the system capacity performance can reach 15 UEs. However, when $M \geq 100$, the system capacity reaches saturated with near optimal solution with 18 UEs. From Fig. 2, it is observed that the large population size can reach the optimal solution.

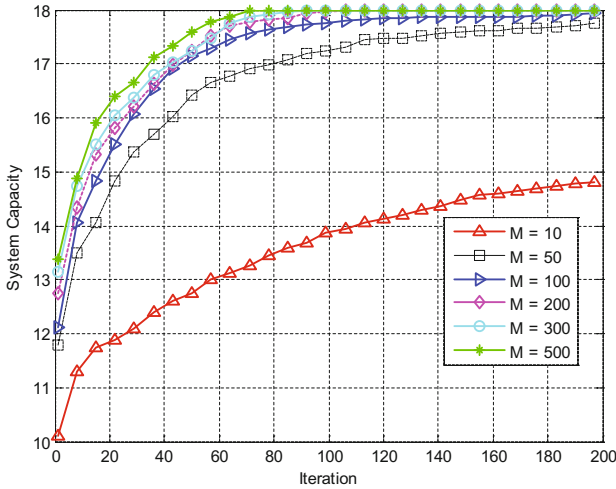


Fig. 2. System capacity for various population size with $K = 18$.

Table 2 shows the performance comparisons of RPSO, GA and Rand methods with $M = 100$ and $G = 200$. From the results in Table 2, it is observed that when $M = 100$, the proposed GA outperforms RPSO with 1.4 UEs of system capacity for $K = 24$.

Table 2. System capacity comparisons of RPSO, GA and Rand methods with $M = 100$ and $G = 200$.

K	12	18	24
GA	12	17.96	21.8
RPSO	12	17.86	20.41
Rand	12	15.08	16.27

5 Conclusions

In this paper, the GA is proposed to optimize resource allocation in D2D systems. Simulation results show that the proposed GA can improve the system capacity performance. With population size $M = 100$ and generations $g = 200$, the proposed GA can outperform the previous proposed RPSO 1.4 users for system capacity.

Acknowledgments. This work was funded in part by Ministry of Science and Technology of Taiwan under Grant MOST 105-2221-E-324-019 and 106-2221-E-324-020.

References

1. Yun, L., Le, Z., Xin, T., Bin, C.: An advanced spectrum allocation algorithm for the across-cell D2D communication in LTE network with higher throughput. *China Commun.* **13**, 30–37 (2016)
2. Camps-Mur, D., Garcia-Saavedra, A., Serrano, P.: Device-to-device communications with wi-fi direct: overview and experimentation. *IEEE Wirel. Commun.* **20**(3), 96–104 (2013)
3. 3GPP - The Mobile Broadband Standard (Rel. 12). <http://www.3gpp.org/>
4. Hasan, M., Hossain, E.: Resource allocation for network-integrated device-to-device communications using smart relays. In: *IEEE Globecom Workshops*, pp. 591–596, December 2013
5. Babun, L.: Extended coverage for public safety and critical communications using multi-hop and D2D communications. Master thesis, Department of Electrical Engineering, Florida International University, March 2015
6. Wang, L., Peng, T., Yang, Y., Wang, W.: Interference constrained D2D communication with relay underlying cellular networks. In: *IEEE Vehicular Technology Conference*, pp. 1–5, September 2013
7. Huang, Y.-F., Tan, T.-H., Chen, B.-A., Liu, S.-H., Chen, Y.-F.: Performance of resource allocation in device-to-device communication systems based on particle swarm optimization. In: *Proceedings of 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC2017)*, Banff, Canada, 5–8 October 2017
8. Goldberg, D.E.: *Genetic Algorithm in Search, Optimization, and Machine Learning*. Addison-Wesley, Boston (1989)
9. Tan, T.-H., Huang, Y.-F., Liu, F.-T.: Multi-user detection in DS-CDMA systems using a genetic algorithm with redundancy saving strategy. *Int. J. Innovative Comput. Inf. Control (IJICIC)* **6**(8), 3347–3364 (2010)
10. Hasan, M., Hossain, E.: Distributed resource allocation for relay-aided device-to-device communication: a message passing approach. *IEEE Trans. Wirel. Commun.* **13**(11), 6326–6341 (2014)
11. Tang, K.S., Man, K.F., Kwong, S., He, Q.: Genetic algorithms and their applications. *IEEE Sign. Process. Mag.* **13**, 22–37 (1996)

Data Mining and Its Applications

Mining Erasable Itemsets Using Bitmap Representation

Wei-Ming Huang¹, Tzung-Pei Hong^{1,3(✉)}, Guo-Cheng Lan²,
Ming-Chao Chiang¹, and Jerry Chun-Wei Lin⁴

¹ Department of Computer Science and Engineering,
National Sun Yat-Sen University, Kaohsiung 804, Taiwan
granthill168@gmail.com, mcchiang@cse.nsysu.edu.tw

² Education Solutions and Service (ESS) Department,
Delta Electronics, Inc., Taipei 114, Taiwan
rrfoheiy@gmail.com

³ Department of Computer Science and Information Engineering,
National University of Kaohsiung, Kaohsiung 811, Taiwan
tphong@nuk.edu.tw

⁴ School of Computer Science and Technology, Harbin Institute of Technology
Shenzhen Graduate School, Shenzhen University, Town Xili, Shenzhen, China
jerrylin@ieee.org

Abstract. This paper proposes a bitmap representation approach for modifying the erasable-itemset mining algorithm to increase its efficiency. The proposed approach uses the bitmap concept to save processing time. Through experimental evaluation, simulation datasets were used to compare the traditional erasable itemset mining and the proposed approach under various experimental conditions.

1 Introduction

Data mining has fascinated a lot of attention, which is the opinion of discovering valuable technology for extracting useful knowledge of pattern in large databases. The relationship among products such that the existence of certain products tends to imply the existence of other products in retail market was studied by mining frequent itemsets and association rules [1, 2]. Frequent patterns has become as an essential problem in data mining and expend many new pattern mining approaches which are derived from [1, 2], such as utility mining [4–7] fuzzy utility mining [11–13] FP-tree [14, 15] erasable itemset mining [3, 8–10].

In 2009, Deng et al. presented a new issue, which is known as erasable itemset mining [8] for analyzing production planning. Assume a factory, which wants to produce many kinds of products. Each product is made up of some materials. Each product can bring in revenue of the factory and each material has a cost of purchase and storage space. To manufacturing products, the manager of factory must spend much money to buy many materials by satisfying the demands of each customer. While the manager of factory has not enough money in economic depression or financial shortage, the insufficient materials may not accomplish productive needs. However, it causes

that the revenue of the factory may be dropped. The spirit of erasable mining is thus to explore the patterns that can be removed, which provide a manager to reduce the revenue of factory under control. Although many methods based on erasable mining [3, 8–10] have been developed, a considerable amount of execution time still remain to be solved. To solve problems mentioned above, we determine that creating a new framework for finding erasable itemsets and overcoming the processing time.

2 Related Works

2.1 Erasable Itemset Mining

Deng et al. [8] proposed the concept of erasable itemset mining which is another deformation of frequent itemset mining. Especially economic depression or financial shortage, many factories want to save money to survive. Consequently, they begin to think how to plan manufacture of products and control stocks of raw materials. The product is made up of materials. How to effectively keep stocks of raw materials is one of the key points. However, reducing stocks of raw materials may cause to stop the manufacture of some products, make it drop down the income of the factory. The erasable itemset mining thus can solve this problem.

In recent years, there were several algorithms are derived from the erasable itemset mining [8], such as VME [10] which uses a new idea, PID_list structure, to store the profit of all products improve the process time and prune the irrelevant data. MERIT [9] which uses a new tree structure, NC_set, keeps overall information to prune unrelated data. WEP [3] provide weighted erasable patterns by considering the different weight of each material to mine user's need.

2.2 Bitmap Representation

A bitmap representation for erasable itemset mining represents as a simple index value (0 or 1). Assume that $A = \{0, 1, 0, 1, 0, 0\}$ and $B = \{0, 0, 0, 0, 1, 1\}$ which include one row and six columns. The bit matrix of A and B include integer value ranging from 0 to 1. For example, we firstly perform logical inclusive *OR* operation between A and B. It can be seen that A and B can be represented as $\{0, 1, 0, 1, 0, 0\}$ and $\{0, 0, 0, 0, 1, 1\}$, and perform logical inclusive *OR* operation to $A \text{ OR } B$, which is $\{0, 1, 0, 1, 1, 1\}$.

3 Problems and Statement

To explain the proposed approach, assume a product database (*PD*) is set in Table 1. Product database $PD = \{P_1, P_2, \dots, P_y, \dots, P_z\}$, where P_y is the y -th product in *PD*. The $M = \{m_1, m_2, \dots, m_i, \dots, m_n\}$ be a set of distinct materials of products in *PD*, where m_i represents the i -th material of product. The profit value, $Profit_i$, is the profit of product P_i in *PD*. An itemset X is a collection of one or more materials; namely, if $|X| = 2$, the itemset X can be treated as a 2-itemset.

Table 1. The example of eleven products in product database.

<i>PID</i>	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>F</i>	<i>G</i>	<i>H</i>	<i>Profit</i>
P_1	✓	✓	✓						2100
P_2	✓	✓							1000
P_3	✓		✓						1000
P_4		✓	✓		✓				150
P_5		✓			✓				50
P_6			✓		✓				100
P_7			✓	✓	✓	✓	✓		200
P_8				✓	✓	✓		✓	100
P_9				✓		✓			50
P_{10}		✓				✓		✓	150
P_{11}			✓			✓			100

The goal of the erasable-itemset mining [8] is to mine unimportant itemsets. For achieving this purpose, some terms for deriving erasable itemsets are introduced as follows.

The gain value of $gain(X) = \sum_{X \in P_y} Profit_y$. Assume $X = \{A, B\}$. From Table 1, the six products P_1, P_2, P_3, P_4, P_5 and P_{10} include sub-1-itemset A or B . Therefore, $gain(X)$ is 4450. Assign a pre-defined minimum erasable itemset mining threshold ratio t , let $T = \sum_{P_y \in PD} Profit_y$. An itemset X , which satisfies $gain(X) \leq T \times t$, is called an erasable itemset. From Table 1, let $t = 35\%$. The item F appears in P_7, P_8, P_9, P_{10} and P_{11} . Its gain is calculated as 600. Item F is called an erasable itemset ($\leq 5000 \times 35\%$). This work presents a new issue research by using bitmap representation approach for mining erasable itemsets can be described as follows:

An erasable X -itemset and $|X| = 1$, which satisfies $gain(X) \leq T \times r$. A bitmap of erasable X -itemset is one-dimensional table, $Bitmap_X = \{b_1^X, b_2^X, \dots, b_j^X, \dots, b_z^X\}$, where z represents the number of products and the value of each b_j represent 1 if sub 1-itemset of X -itemset at least exists in P_j Product, otherwise 0. For example, the $Bitmap_A = \{1, 1, 1, 0, 0, 0, 0, 0, 0, 0\}$. The erasable X -itemset can be decomposed $|X|$ 1-itemsets and $|X| > 2$. That is, the bitmap of erasable X -itemset is $Bitmap_X = \{b_1^X, b_2^X, \dots, b_n^X\} = \{b_1^1 OR b_1^2 OR \dots OR b_1^{|X|}, b_2^1 OR b_2^2 OR \dots OR b_2^{|X|}, \dots, b_z^1 OR b_z^2 OR \dots OR b_z^{|X|}\}$, where $Bitmap_X$ is the bitmap of the X -itemset, $|X|$ is the number of sub 1-itemset of X -itemset, $OR b_z^{|X|}$ represents whether $|X|$ -th 1-itemset of X -itemset at least exists in P_z Product or not, in other that each pair of corresponding bits carry out logical inclusive OR operation using two bit patterns of equal length. For example, the $bitmap_A$ and the $bitmap_B$ can be represented as $\{1, 1, 1, 0, 0, 0, 0, 0, 0, 0\}$ and $\{1, 1, 0, 1, 1, 0, 0, 0, 0, 1, 0\}$, and perform logical inclusive OR operation to the $bitmap_{\{A, B\}}$, which is

{1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 1, 0}. The gain of erasable X -itemset is the summation of $profit_j$ value of bit value b_j on $Bitmap_X$. For example, the $bitmap_{\{A, B\}}$ is {1, 1, 1, 1, 1, 0, 0, 0, 0, 1, 0}. The $gain(\{A, B\}) = 4450$.

4 The Bitmap Representation for Erasable Mining Approach

The *BERM* method is proposed to mine erasable itemsets for the bitmap representation. The pseudo-code of the *BERM* algorithm is given in Fig. 1. The *BERM* method is two main subroutines: (1) to sum the profit of one product to the gain value of the material in product database if the material exists in one product and then build the bitmap value. If the gain value is smaller than or equal to threshold value, it is an erasable itemset (Fig. 2). (2) to find erasable $(r + 1)$ -itemset from erasable r -itemset bitmap which we get and then build the bitmap value of erasable $(r + 1)$ -itemset (Fig. 3).

Algorithm. BERM Algorithm.

Input: PD , a product database; n products identification $PD = \{P_1, P_2, \dots, P_n\}$; the profit of the products with $\{profit_1, Profit_2, \dots, Profit_n\}$ values; all products produced with $\{m_1, m_2, \dots, m_n\}$ materials; a pre-defined minimum erasable itemset mining threshold t .

Output: A final set of erasable itemsets (EIs).

```

// T is the sum up the profit of the products in PD
1. FOR each product  $P_y \in PD$  DO
2.    $T = T + P_y.profit_y$ ;
3. END FOR
// gain erasable 1-itemset and build bitmap value
4.  $E_1 = \text{gain } E_1 \text{ itemset and construct } E_1 \text{ bitmap}$ ;
5.  $EI = EI + E_1$ ;
//gain erasable (r+1)-itemset from erasable r-itemset bitmap
6.  $set\ r = 1$ ;
7. WHILE ( $E_r \neq \emptyset$ ) DO
8.    $E_{r+1} = \text{gain } E_{r+1} \text{ itemset from } E_r \text{ itemset bitmap}$ ;
9.   IF  $E_{r+1} \neq \emptyset$  THEN
10.     $EI = EI + E_{r+1}$ ;
11.   ELSE
12.    go to line 16;
13.   END IF
14.    $r = r + 1$ ;
15. END WHILE
16. RETURN  $EI$ ;
```

Fig. 1. BERM algorithm

Procedure 1. gain E_1 itemset and construct E_1 Bitmap.

Parameters :

T is the sum up the profit of the product in PD ,
a pre-defined minimum erasable itemset mining threshold t

Output: A set of erasable 1-itemsets (E_1).

```

/* scan database to find erasable 1-itemset and construct
  bitmap value*/
1. Threshold = T * t;
2. FOR each material  $m_i$  in  $PD$  DO
3.   FOR each product  $p_y$  in  $PD$  DO
4.     Bitmap $_i$  = {0, 0, ..., 0} // = {b $^1_1$ , b $^1_2$ , ..., b $^1_y$ }
5.     IF  $m_i \in p_y$  THEN
6.       gain( $m_i$ ) = gain( $m_i$ ) +  $P_y$ .profit $_y$ ;
7.       b $^1_y$  = 1;
8.     END IF
9.   END FOR
10.  IF gain( $m_i$ ) <= Threshold THEN
11.     $E_j = E_j + \{m_i\}$ ;
12.  END IF
13. END FOR

14. RETURN  $E_j$ ;

```

Fig. 2. Gain E_1 itemset and construct E_1 bitmap for procedure 1.

Procedure 2. gain E_{r+1} itemset from E_r itemset bitmap.

Parameters :

T is the sum up the profit of the product in PD ,
a pre-defined minimum erasable itemset mining threshold t
 E_r itemset Bitmap value

Output: A set of erasable $(r+1)$ -itemsets (E_{r+1}).

```

1. WHILE  $E_r \neq \emptyset$  DO
2.    $CE_{r+1} = \text{Apriori}(E_r)$ ;
3.   Bitmap $_{r+1}$  = {0, 0, ..., 0}; // = {b $^{r+1}_1$ , b $^{r+1}_2$ , ..., b $^{r+1}_y$ }
4.   FOR each  $(r+1)$ -itemset  $X$  in  $CE_{r+1}$  DO
5.     FOR each 1-sub-itemset  $i$  in  $X$  DO
6.       Bitmap $^X_{r+1} = \text{Bitmap}^X_{r+1} \cup \text{Bitmap}^X_i$ ;
7.     END FOR
8.   END FOR

9.   FOR each  $(r+1)$ -itemset  $X$  in  $CE_{r+1}$  DO
10.    gain $^X_{r+1} = 0$ ;
11.    FOR each product  $p_y$  in  $PD$  DO
12.      IF b $^{r+1}_y = 1$  THEN
13.        gain $^X_{r+1} = \text{gain}^X_{r+1} + P_y$ .profit $_y$ ;
14.      END IF
15.    END FOR
16.    IF gain $^X_{r+1} <= \text{Threshold}$  THEN
17.       $E_{r+1} = E_{r+1} + \{X\}$ ;
18.    END IF
19.  END FOR
20. END WHILE

21. RETURN  $E_{r+1}$ ;

```

Fig. 3. Gain E_{r+1} itemset from E_r itemset bitmap for procedure 2.

5 Performance Evaluation

The proposed method *BERM* and *META* [8] were implemented in J2SDK 1.8 on a computer with an Intel Core i5-4590 CPU 3.3 GHz and 16 GB RAM. They were evaluated on synthetic dataset, which are based on the IBM data generator [16]. The profit value of each product was generated between normal distribution $N(100, 500)$. The threshold is set at between 10% and 50%. We compare the performance of *BERM* and *META* in Fig. 4.

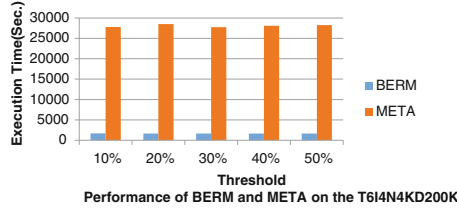


Fig. 4. Execution time of T6I4N4KD200 K datasets

6 Conclusion

An algorithm for the bitmap concept was proposed, in this paper, to determine the erasable k -itemsets based on $(k-1)$ -itemsets. Bitmap representation approach, then, were defined for quickly mining erasable itemsets information. To point out the efficiency of *BERM*, some experiments are conducted to compare *BERM* and *META* regarding execution time for T6I4N4KD200K dataset. The experimental results show that *BERM* outperforms *META* in execution time of mining.

References

1. Agrawal, R., Srikant, R.: Fast algorithm for mining association rules. In: The 20th International Conference on Very Large Data Bases, pp. 487–499 (1994)
2. Agrawal, R., Imieliński, T., Swami, A.: Mining association rules between sets of items in large database. In: The ACM SIGMOD International Conference on Management of Data, pp. 207–216 (1993)
3. Lee, G., Yun, U., Ryang, H.: Mining weighted wrasable patterns by using underestimated constraint-based pruning technique. *J. Intell. Fuzzy Syst.* **28**(3), 1145–1157 (2015)
4. Chan, R., Yang, Q., Shen, Y.D.: Mining high utility itemsets. In: The 3rd IEEE International Conference on Data Mining, pp. 19–26 (2003)
5. Yen, S.J., Lee, Y.S.: Mining high utility quantitative association rules. In: The 9th International Conference on Data Warehousing and Knowledge Discovery, pp. 283–292 (2007)
6. Liu, Y., Liao, W.K., Choudhary, A.: A fast high utility itemsets mining algorithm. In: The 1st International Workshop on Utility-Based Data Mining, pp. 90–99 (2005)

7. Liu, Y., Li, J., Liao, W.K., Choudhary, A., Shi, Y.: High utility itemsets mining. *Int. J. Inf. Technol. Decis. Mak.* **09**(06), 905–934 (2010)
8. Deng, Z.H., Fang, G.D., Wang, Z.H., Xu, X.R.: Mining erasable itemsets. In: *The 8th International Conference on Machine Learning and Cybernetics*, pp. 67–73 (2009)
9. Deng, Z.H., Xu, X.R.: Fast mining erasable itemsets using nc_sets. *Expert Syst. Appl.* **39**(4), 4453–4463 (2012)
10. Deng, Z.H., Xu, X.R.: An efficient algorithm for mining erasable itemsets. In: *The 6th International Conference on Advanced Data Mining and Applications*, pp. 214–225 (2010)
11. Lan, G.C., Hong, T.P., Lin, Y.H., Wang, S.L.: Fuzzy utility mining with upper-bound measure. *Appl. Soft Comput.* **30**, 767–777 (2015)
12. Chen, C.H., Li, A.F., Lee, Y.C.: Actionable high-coherent-utility fuzzy itemset mining. *Soft. Comput.* **18**(12), 2413–2424 (2014)
13. Wang, C.M., Chen, S.H., Huang, Y.F.: A fuzzy approach for mining high utility quantitative itemset. In: *The IEEE International Conference on Fuzzy Systems*, pp. 1909–1913 (2009)
14. Lin, C.W., Hong, T.P.: A new mining approach for uncertain databases using CUFPTrees. *Expert Syst. Appl.* **39**(4), 4084–4093 (2012)
15. Hong, T.P., Lin, C.W., Lin, T.C., Chen, Y.F., Pan, S.T.: Integration of multiple fuzzy FP-trees. In: *Intelligent Information and Database Systems*, pp. 330–337 (2012)
16. IBM Quest Data Mining Projection, Quest synthetic data generation code (1996). <http://www.almaden.ibm.com/cs/quest/syndata.htm>

Identifying Suspicious Cases in the Hong Kong Stock Market Using Commentators' Stock News

Li Quan^(✉) and Jean Lai

Hong Kong Baptist University, Kowloon, Hong Kong
quanlime@gmail.com, jeanlai@comp.hkbu.edu.hk

Abstract. Stock market is one of the most active secondary markets in the finance industry. Investors buy and sell stocks there. Regulators are in place to offer a fair place for trading. Unfortunately, market manipulation of different forms still exists. Limited researches worked on this area due to restriction in data supply. Exchanges and regulators do not offer full order book to the public. Without a full order book, it is difficult to locate the fraud intention of manipulator. This paper introduces a new method to identify suspicious cases: identifying suspicious cases by parsing the stock recommendation news written by stock commentators. Some suspicious cases are found in the study.

Keywords: Stock price manipulation · Suspicious case · Stock news analysis · Correlation coefficient

1 Introduction

People are willing to invest their money in stock market. Rational investors trade to profit from the market. Manipulators, one kind of market participants, intend to manipulate stock prices so that they can profit in short. Ordinary investors like to trade active stocks due to its liquidity. Manipulators therefore intend to produce such a misconception. They ramp up and down the stock price with a means which is hard to be captured by the regulator. Obviously manipulating the stock price with their own money is not a wise choice. The most common means to manipulate a stock is to buy a stock when its price is still low, and then attract ordinary investors to buy, projecting a high volume. Economic theory told us that the stock price would grow up due to high demand then. Manipulators would realize the profit by selling the stocks when the stock is pushed to a high price.

Detecting stock market manipulation is difficult as full order book is not easily to be accessed. Exchanges and regulators provide part of the order book to outsiders only. This study therefore aims to identify a new method to detect stock price manipulation cases. Stock commentators often recommend stocks through media like TV channel, newspapers, and etc. Ordinary investors without doing analysis on their own belief in commentators' recommendations very much. As such, some commentators might take this as an advantage for their only benefit. They are required to declare their holding in the recommended stocks when doing recommendation in media, but it seems to have

no legal restriction not allowing them to recommend the stocks which are on their hand or whatever. This study is to examine the correlations between pre-recommended and post-recommended stock prices and volumes, and study its relationship between the movements and the actual performance of this stock fundamentally and technically. Since news are all in text, so a dictionary-based method is used to recognize the commentators' recommendations.

2 Literature Review

2.1 Stock Market

Stock market is a platform for trading equities, derivatives and other securities, either through exchanges or over-the-counter markets. Stock market, is also known as equity market, which is one of the most vital components of a free-market economy. Also, it provides firms with access to capital in exchange for giving investors a slice of ownership [1].

2.2 Market Manipulation

Manipulation of stock price refers to some stock investors decided to obtain huge profits, by controlling the stock investment information of other investors having reference to, control the future trend of stock price behavior. Manipulation of stock price is not necessarily done by the organizational investors, but also individual investors.

2.3 Price Manipulation Detection

Many exchanges and regulators such as the Securities and Futures Commission (SFC) of Hong Kong use rule-based system to detect suspicious cases. This program generates over thousands of alerts daily. All suspicious cases would be reviewed by the staff there manually. Only few cases will be further investigated which is not so efficient. Some detection methods which are based on mathematical models such as the pump-and-dump model and spoof trading model are also available to detect price manipulation [3]. Effectiveness and efficiency in the existing detection mechanisms can be further improved.

3 Data Preparation

3.1 Data Source

We use WiseNews as data source. WiseNews provides full-text articles from over 600 key newspapers and magazines, Websites from China, Hong Kong, Macau, Taiwan, and the United States. It includes news up to today. It has Hong Kong newspapers from late 1998 to current, China, Macau, Taiwan newspapers from 2000 to current, and U.S. regional newspapers from 2000 to current. Moreover, over 45,000 sources of news and information covering print, web, multimedia and more.

The attributes of our dataset include heading, author, newspaper name, content and news ID.

3.2 Data Extraction

Data were extracted using a program written in Node.js. Node.js is a JavaScript runtime built on Chrome's V8 JavaScript engine [4]. Node.js uses an event-driven, non-blocking I/O model that makes it lightweight and efficient. So, we choose this language as solution to extract news.

Web scraping is data scraping used for extracting data from websites. Web scraping software may access the World Wide Web directly using the Hypertext Transfer Protocol, or through a web browser [5]. We implement a multi-process scraper. Many workers can run simultaneously which can boost the efficiency of scraping.

After scraping webpage from wise news, we get sufficient data about 54000 records in one year period.

3.3 Data Formatting

What we want to learn is stock. So, the most important thing is to identify stock number from the news paragraph. From observation, we found that all the stock ids are contained between a pair of brackets like “滙豐控股 (00005)、友邦保險 (01299)”. One news often mention more than one stocks, which means stock id may appear in every sentence of the news. Sometime, the stock news does not contain any stock id. Stock id is then retrieved with the stock name. We also removed all the irrelevant sentences, that are the sentences do not contain stock id. Finally, the result of the formatted dataset is that every sentence which contains stock id will be extracted.

Around 67,000 sentences are having a stock id. With the stock id, we can refer to the stock's closing price and trading volume. Moreover, the trend of stock price we want to capture is that the stock price will change after this news was published. We called the news' publish day as t day. The next trading day as $t + 1$ day, we compute the increase ratios in following $t + 5$ day, the formula of computing increase ratios is:

$$increase\ ratio_{t+i} = \frac{price_{t+i} - price_t}{price_t} \quad (1)$$

We compute increase ratios from t to $t + 5$, where $i = 1, 2, 3, 4,$ and 5 . When it comes to the volume, we assume that the change of volume should occur in t day, so we must compare the past day like $t - 1$ day to t day volume. Also, we compute the mean of $t - 5$ to $t - 1$ volume, and use this formula to calculate increase ratio of volume.

$$increase\ ratio_{volume} = \frac{Volume_t - Mean}{Mean} \quad (2)$$

This formula shows the comparison ratio between average level of volume in past 5 days and current day.

3.4 Dictionary Building

In order to evaluate the news recommendation, a dictionary with the polarity of the Chinese words in the news was constructed. We assume that the commentators' attitude toward this stock can be seen by the polarity of the news. We conclude some words that usually used to describe the stock. Some examples were shown in Table 1.

Table 1. Data dictionary

Positive statement (27 words)	Negative statement (18 words)
起跑,回升,反彈,潛力十足,利好,贏,推介,增長,受惠,有望,造好,漲,跟進,做好,憧憬,驚喜,吸資,非凡,搶眼,復甦,升,突破,向淡,增,上望,升幅,回穩	跌,錯,減,低,瀉,挫,蝕,弱勢,插,蒸發,吐,做淡,屠牛,下滑,流走,軟,回落,下降

We then count the words of positive and negative and store the number in attribute of rise and drop.

Finally, in this new dataset, it is consisted of attributes of descriptions for storing sentences, stocks for storing stock id, time for storing time of the news being published, rise for counting number of rising statements appearing in sentence, drop for counting number of drop statements appearing in sentence, and price increase rate from t to $t + 5$ and volume increase rate.

4 Data Analysis and Result

4.1 Assumptions

Firstly, we should set some assumptions about how to identify suspicious cases.

Level of Recommendation. We extracted the sentences which included the stock id only and counted the number positive and negative words in the sentences to determine the level of recommendation. If the commentator uses more positive words to describe a stock, implying that this commentator “recommend” this stock to readers.

Volume Change. Market efficiency theory stated that the demand of a stock would grow up or down naturally with information that are available to all market participants. Participants react to positive or negative news quite consistently, resulting in ramping up or down in volume heavily. Volume change is one of the most obvious indicators to figure out the demand of a stock. If the news is supported by some fundamental facts, such as increase in annual profit, the volume growth is said to be normal and natural. Otherwise we may conclude this change as usual. Unusual change in volume can be regarded suspicious.

Volume manipulation is one kind of manipulation forms that we often see in exchanges. In a court case, manipulator wanted to create a false or misleading appearance of trading and share price movements in Bauhaus shares that had misled

other investors that there are many investors to invest this stock. As a result, he made a profit of \$25,492 [6].

Price Change. Most of time, it is common to see fluctuation of stock price, which means we cannot judge suspicious cases from price change independently. We should consider the volume change of the stock at the same time. For example, there is a big increase in volume while the price increases next day. Again, if we do not see any fundamental facts to support the increase, we can conclude this increase as suspicious as well.

4.2 Correlation Coefficient

Pearson’s correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. Also, it is a measure of the linear correlation between two variables X and Y [7].

Positive Level versus Increase Rate of Price. We compute the Pearson Partial Correlation Coefficient with variable of increase rate of price and number of positive words. In Table 2, it shows some cases that the coefficient of increase rate $t + 1$ larger than 0.4. In this table, we get 5 coefficients in rows. Some of the commentators’ coefficients are relatively higher than the rest. It means that if these commentators say a specific stock will increase, then its stock price would increase in the following five days, especially the first day. It is wired for stock commutators to predict so accurately. Maybe it is accidental or not, we will keep explore in following part. This table will be used as reference table.

Table 2. Partial correlation in stock price change in the five days after the news publish day (by Commentator)

Commentator(s)	Increase Rate T+1	Increase Rate T+2	Increase Rate T+3	Increase Rate T+4	Increase Rate T+5
柯慧琳 林衛	0.624	0.332	0.142	-0.207	-0.118
柯慧琳 林衛 鄧凱玲	0.612	Not Significant	0.758	0.866	0.780
周顯	0.551	0.033	-0.123	-0.053	-0.020
羅偉文	0.407	0.362	0.424	0.485	0.482

Polarity of the Commentator’s News versus Change in Stock Volume. In Table 3, we compute the correlation coefficient between the polarity of the commentator’s news and the change in t-day volume. We found the correlation in volume change is much higher than the increase rate in price generally. This illustrates the detection of suspicious cases by volume change is more significant than by stock price.

Table 3. Partial correlation in t-day volume (by Commentator)

Commentator	Change in t-Day Volume
譚曉涵	0.790
鄭福發	0.718
陳韻妍何仲豪	0.590
周文婷楊玉燕	0.568
周顯	0.566
...	...

4.3 Ratio of Exceed Average Volume

After getting two reference tables, we can join two tables. Then, we pick up volume increase ratio which is larger than 100%. It means that stock which is mentioned on the stock news had change in volume after having this news. It is possible that people start to buy this stock after reading this news. That is why volume of the stock changes significantly. In Table 4, all the increase ratio is larger than 100%. Most of the stocks are having two times or even twelve times increase in the volume in t-day.

Table 4. Increase ratio of stocks

Stock	Publisher	Time	Rise	Drop	Author	Increase Volume Ratio
01340. HK	蘋果日報	2016-06-22	2	2	周顯	1.84
00601. HK	蘋果日報	2016-09-28	0	0	周顯	8.72
00261. HK	明報	2016-12-07	2	2	周顯	1.13
01053. HK	投資理財	2016-10-11	1	0	周文婷楊玉燕	1.70
...

However, we cannot draw a conclusion to prosecute these commentators though they have highly correlation coefficient to stock volume and price. The finally step is to identify supporting clues by the regulator.

4.4 Supporting Clues

Briefly, we start to find related news to boost the price or the volume of the stock in this part. Any news which can support commentators' opinions will be considered as objective clue. For example, a firm's financial statement went public. Once there are no clues to support commentators' opinions, the coefficient is so high. This news will then be regarded as suspicious cases. We search support news in HKEx News [8] for demonstration. Take 1340.HK as an example shown in Table 5.

Table 5. A sample suspicious case

Stock	Publisher	Time	Rise	Drop	Author	Average Volume	Increase Volume ratio
01340.HK	蘋果日報	2016- 06-22	2	2	周顯	7630400	1.84

In Table 6, we will find the increase ratio of volume is 184%. That is, almost double the original average volume based on past five trade days. According to materials we found in the HKEx News. This firm will outstand new portion of shares, which makes the trend of volume of the stock fluctuate unusually which sounds normal.

4.5 Result

After processing step by step, we get a suspicious cases table, Table 6 shows the possible suspicious cases. The suspicious cases do not have any fundamental factors

Table 6. Suspicious cases

Stock	Publisher	Time	Rise	Drop	Author	Average Volume	Increase Volume ratio
00601.HK	蘋果日報	2016- 09-28	0	0	周顯	2005125.2	8.72
00261.HK	明報	2016- 12-07	2	2	周顯	399397420	1.13
02318.HK	投資理財	2016- 10-11	2	0	周文婷楊 玉燕	17985620.7 5	1.08
02016.HK	信報財經新聞	2016- 05-18	0	0	譚曉涵	64000	2.16
02888.HK	投資理財	2016- 03-21	0	0	李永權	2354455.25	2.40

(such as supporting news or another clue) to support the sudden increase in volume. We have sufficient confidence to say these commentators have relatively high possibility to “manipulate” stock price or their recommendation is quite convincing to many market participants.

We can see many increase volume ratios of many cases are much bigger than 1. Even 10 times larger than common average volume. Many people invest their money on these stocks after reading the famous commentators’ opinions.

5 Conclusion

This study has some limitations. The first is that the study is based on stock news written by commentators. We can only identify some suspicious cases and the total number of cases will be lower than that generated by the existing approach. We however cannot conclude the cases as manipulation cases due to insufficient fraud evidence. In addition, the case extraction cannot be performed in real-time basis yet.

The other is evaluation of recommendation level. We have sufficient reason to show count word is uncomplicated way. But how to judge the polarity of the sentence should be more technical. Many advanced algorithms like Bayesian Network and Neural Network should be applied in future study.

References

1. What is stock market. <http://www.investopedia.com/terms/s/stockmarket.asp>
2. Investor convicted of market manipulation. <http://sc.sfc.hk/gb/www.sfc.hk/edistributionWeb/gateway/EN/news-and-announcements/news/doc?refNo=13PR76>
3. Leangarun, T., Tangamchit, P., Thajchayapong, S.: Stock price manipulation detection based on mathematical models. *Int. J. Trade Econ. Finance* 7(3), 81–88 (2016). <http://dx.doi.org/10.18178/ijtef.2016.7.3.503>
4. Node.JS Official Website. <https://nodejs.org/en/>
5. What is web scraping? https://en.wikipedia.org/wiki/Web_scraping
6. Retail investor fined for market manipulation
7. <http://www.sfc.hk/edistributionWeb/gateway/EN/news-and-announcements/news/doc?refNo=10PR43>
8. Pearson correlation coefficient. https://en.wikipedia.org/wiki/Pearson_correlation_coefficient#Definition
9. HKEx News. http://www.hkexnews.hk/index_c.htm

A New Conceptual Model for Big Data Analysis

Thi Thi Zin¹(✉), Pyke Tin¹, and Hiromitsu Hama²

¹ University of Miyazaki, Miyazaki, Japan
thithi@cc.miyazaki-u.ac.jp, pyketinll@gmail.com

² Osaka City University, Osaka, Japan
hama@ado.osaka-cu.ac.jp

Abstract. In today modern societies, everywhere has to deal in one way or another with Big Data. Academicians, researchers, industrialists and many others have developed and still developing variety of methods, approaches and solutions for such big in volume, fast in velocity, versatile in variety and value in vicinity known as Big Data problems. However much has to be done concerning with Big Data analysis. Therefore, in this paper we propose a new concept named as Big Data Reservoir which can be interpreted as Ocean in which all most all information is stored, transmitted, communicated and extracted to utilize in our daily life. As a starting point of our proposed new concept, in this paper we shall consider a stochastic model for input/output analysis of Big Data by using Water Storage Reservoir Model in the real world. Specifically, we shall investigate the Big Data information processing in terms of stochastic model in the theory of water storage or dam theory. Finally, we shall present some illustrations with simulation.

Keywords: Big Data · Stochastic model · Water Storage Reservoir Model

1 Introduction

From industry to consumer, banking to retail, from medical expert to patient and many other sectors have already been and have being embraced into Big Data - regardless of whether the information come from private or public. In terms volume, the scaling of petabyte data has been flowed into big data reservoir daily from web services, social media, astronomy, and biology science. Thus, big data can be defined as a collection of large datasets that may not be processed using traditional database management tools [1]. Mostly, problems of data storage, information manipulation, and especially searching key information from big data have been front line research areas which are widely researched and engineered by sizable researchers [2]. Specifically, the data flow into big data has two sources such as collective gathering and individual generation. The collective gathering big data includes smart city data, national geographic conditions monitoring data, and earth observation data [3]. Usually the collective gathering data are obtained by using statistical sampling techniques leading to high quality data. On the other hand, most of big data are generated by individuals by using social media data on the Internet. The individual data generation is more freedom giving low reliability and usability [4].

Collecting and analyzing data are commonly concerned with statistics which make able to judge on the basis of hypothesis. However, the big data technology is much advanced at the stages of data sampling, storage management, data computation, and data communication. In the traditional scientific paradigm, the theory is proved with the experiments. In the current scientific paradigm, the scientific finding is often obtained by computer simulation, and is mainly explored from multi-source observations from big data. In summary, we can say that big data have the characteristics of 4 V namely, volume, variety, velocity, and veracity. As the names describe the meaning we can explain these characteristics as follows. First V stands for volume of large amount of data and the second V represents variety or multiple-type or multi-source data. The third V representing velocity of generating data and processing data at high speed and the fourth V characterize veracity or value which refers to the high quality and value of captured and analyzed data. Data quality is comprehensively measured with inherent information content and its user demands satisfaction [5].

In this paper we propose a new concept and approach to the big data by introducing an analogy of big data with reservoir theory in the stochastic water storage processes. We then analyze the data inflows and outflows into buffers to investigate the insight patterns of big data to extract some important key information. The rest part of the paper includes some related works in Sect. 2, the overview and problem formulation in Sect. 3, illustrative simulations results in Sect. 4 and concluding remarks in Sect. 5.

2 Some Related Works

In this section, we shall present some research works of others which are related to this paper. Although a tremendous amount of research works concerning with Big Data analysis and Theory of Storage so called the stochastic reservoir theory has been appeared in the literature, we will describe some works which are in line with our works of this paper. The theory of storage with respect to probability concepts was first introduced by P.A.P Moran in 1954. Since then many researchers had examined and extended the works of Moran in theoretical aspect as well as application aspects. Among them, we would like to refer some works of Phatarfod [5] about stochastic reservoir theory and its extensions [6–9]. In this concern, a common concept is that a reservoir is built for preventing floods or irrigation use in which water inflows into the reservoir and released a certain amount for optimal regulation of a system in which the inflows and outflows are formed a sequences of random variables satisfying laws of probability so that the name becomes a stochastic reservoir theory. Even though we name a stochastic reservoir theory, there are many parts which demand the use of statistical techniques such as time series analysis of inflows to estimate an optimal size of a reservoir. Also, by using statistical regression analysis, we can find the probability distributions of inflows so that the input-output analysis is done for computing various important quantities including storage size, optimal regulation policies, overflow and emptiness probabilities and the respective times to be take. Those quantities are very useful to draw some analogies between the stochastic reservoir theory and buffer data storage system in the Big Data Analysis.

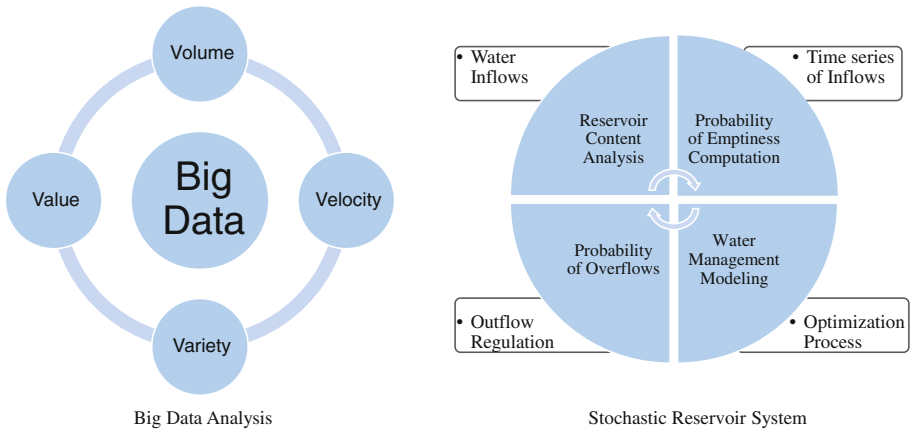


Fig. 1. Stochastic Reservoir and Big Data Models

In the context of Big Data analysis, it has been well recognized that various sources such as social networks generate a huge amount of data having big in volume, fast in velocity, vary in variety and high in values. It will be worthwhile to note concerning with Big Data where the generation process of data is done by the users as well as providers so that most of data are unstructured making the analysis attractable for extracting some reliable and useful information. In order to extract important key information, we make big data processing by establishing some kinds of buffer storage systems where multiple streams of multiple data flow into the buffers of the big data. This is the analogy to be taking into account between the stochastic reservoir theory and big data analysis which is illustrated in Fig. 1.

There is so much innovation in data platforms to enable the efficient processing of new type analytics through novel data structures which is termed as data reservoir which needs a constant flow of new data where ever it comes from existing sources or new sources in order to investigate the usefulness and reliability of the information. The data in the data reservoir can be stored in the data reservoir’s index or catalog. The catalog defines the origin, owner, and the characteristics of the data. Similar to water reservoir, the data reservoir can provide the regulation of inflows and out flow of the data for optimum usages. The data reservoir is designed to offer access the data for analytics [8]. On the other hand, big data reservoirs are large catchment areas where all kinds of data can be stored and analyzed. This fact is the foundation of a natural information systems of future developments. Also, as the data volumes are large, velocity high and the variety huge, it is essential for data reservoirs to be cost effective and flexible to other methods.

Searching data usually provides an automatic and effective way for reservoir evaluation. In [10], the authors considered a technique of data searching for logging reservoir evaluation and verify its performance. Compared with traditional evaluation methods, it is efficient and not so dependent on expertise. In the following we propose a framework which would be able to analyze.

3 Proposed Stochastic Model for Big Data Analysis

In order to make an optimal operation of data reservoir networks for the big data we propose a new concept of stochastic model from water storage systems to analyze big data reservoir networks systems in the big data. In typical models, the inputs to the reservoir are data released from upstream reservoirs and stochastic inflows from external sources (like social media and smart phones), while the outputs correspond to the amount of information to be transmitted during a given time period (e.g., a day or an hour or a month). The dynamics for the single reservoir can be modeled by a state equation where the amount of information at the beginning of period $t + 1$ reflects the flow balance between the information that enters (upstream releases and stochastic inflows) and the information that is released or transmitted during period t . The amount of information to be released during a given time period from each reservoir is chosen to minimize some possibly nonlinear cost (or maximize benefit) function related to the releases. Thus, optimal management of the reservoir network can be formulated as an optimization problem, in which the aim is to determine the quantity of information releases that minimize a total cost over a given horizon of T time periods (e.g., a year).

Let X_t for $t = 0, 1, 2, \dots$ be the amount of data flows into the data reservoir during the time interval $(t, t + 1)$ and at the of interval $t + 1$, a certain amount of data say m is taken out then the data content Z_t left in the reservoir at the end of interval after the release is given the following stochastic state equation:

$$Z_{t+1} = \min(K, Z_t + X_t) - \min(m, Z_t + X_t) \quad (1)$$

The Eq. (1) can be solved when the probability distribution of $\{X_t\}$ and constant m are known. In this paper we shall consider the case of independent and identical distribution for $\{X_t\}$. In order to do so, we first state Wald's fundamental identity in sequential analysis.

Let us define $Y_t = X_t - m$. Then $\{Y_t\}$ is also a sequence of independent and identically distributed random variables. We also have from Eq. (1) that $Z_N = \sum_{t=1}^N Y_t$. In this, there are two absorbing barriers at $-u$ and $K - m (>0)$, for the random walk starting at the origin. Here u is the dam content between 0 and $K - m$.

Let n be the smallest positive integer such that $Z_n \geq K - m - u$ or $Z_n \leq -u + m$. Then, if $G(\theta)$ is the probability generating of the distribution of Y :

$$E [e^{-\theta} Z_n \{G(\theta)\}^{-n}] = 1 \quad (2)$$

for all θ such that $|G(\theta)| \geq 1$.

It is known by Wald that there is one dominant root θ_0 such that $G(\theta) = 1$. Substituting θ_0 in Eq. (2), we can obtain the probability of absorbing at the barrier which is approximately equal to:

$$P_u = \frac{1 - e^{-(K-m+1-u)\theta_0}}{e^{(u+1-m)\theta_0} - e^{-(K-m+1-u)\theta_0}} \quad (3)$$

By assuming the capacity of data reservoir as K and the initial data level as u , with unit release $m = 1$, we then have the probability of data emptiness and flooding data of overflow as described in Eqs. (4) and (5).

$$P_u = \frac{1 - e^{-(K-u)\theta_0}}{e^{u\theta_0} - e^{-(K-u)\theta_0}} \tag{4}$$

$$Q_u = \frac{1 - e^{u\theta_0}}{e^{-(K-u)\theta_0} - e^{u\theta_0}} \tag{5}$$

From the duality theorem of random walk namely as $P_u = F(K - u)$, we obtain the stationary distribution of data reservoir content as shown in Eq. (6).

$$F(x) = \text{Prob (Data content} \leq x) = \frac{e^{x\theta_0} - 1}{e^{K\theta_0} - 1} \tag{6}$$

Now the problem become to evaluate the value of K or the size of data buffer such that:

$$F(K) = \text{Prob (reservoir content} \leq K - 1) = P \text{ where } P \text{ is to be given} \tag{7}$$

4 Experimental Simulation Results

First we made a test how the approximate results given in Eq. (7) are reasonably acceptable or not by comparing the ground truth done by Prabhu [11] for the water content in the theory of storage. The comparison results are presented in Table 1.

Table 1. Distribution function of reservoir content for independent gamma inputs with parameter α and 100 units' release; size of the reservoir, $\sim = 1000$

Reservoir content	$\alpha = 1/0.9$	$\alpha = 1/0.9$	$\alpha = 1/1.8$	$\alpha = 1/1.8$
x	Exact	Approximate	Exact	Approximate
0	0.031	0	0.005	0
100	0.068	0.042	0.016	0.0132
200	0.115	0.094	0.033	0.0323
300	0.172	0.158	0.059	0.0614
400	0.242	0.237	0.097	0.1045
500	0.331	0.333	0.156	0.1716
600	0.435	0.452	0.245	0.2717
700	0.565	0.598	0.379	0.4238
800	0.721	0.778	0.581	0.6529
900	1	1	1	1

From the Table 1, it can be seen that the approximate results are good while the content lies between 200 and 600. That means if the content is between the range of $(K/400, 3K/400)$, the results get better. On the other hand, the results are not very much good when the content is near to zero. This fact does not make much effect for big data because the big data will never become shortage of data. Therefore, we regard that the approximate results for the data reservoir content will make sense for use to determine the reservoir size optimally.

In order to do so, we note that the reservoir size K satisfies the equation $F(K) = P$ specified probability say p in Eq. (6). Putting $e^{lK\theta_0} = z_0$ into Eq. (6),

$$\frac{e^{y\theta_0-1}}{e^{K\theta_0} - 1} = \frac{z_0 - 1}{z_0^{(1/l)} - 1} = P \quad (8)$$

It is known from the solution of a polynomial equation, we can obtain the unique solution of (8) other than non-zero. We then have the optimal reservoir size as:

$$K = \frac{n \log z_0}{\theta_0} \quad (9)$$

By the simulation results we note that for reservoirs with the same values of l and P , $K\theta_0$ is a constant. Thus, to compare sizes of reservoirs with the same critical level and critical probability, we need only compare the values of θ_0 . Thus we can compare the values of θ_0 by varying the size of data reservoirs so that we can obtain the optimal size of data reservoir of computing buffer size in the big data. This will make efficient computing effect to investigate insight information for the big data. However, in this paper we can give the outline procedures only.

5 Conclusion

In this paper we had proposed a new concept of big data reservoir by using the concept of stochastic water storage to investigate the insight information from the big data. We have used Wald's fundamental theorem in sequential analysis which was very popular in various fields of research such as queuing theory, dam theory and other operation research problem. We have presented only simulation so far. More works to be done in the future.

Acknowledgment. This work is partially supported by the Grant of Telecommunication Advanced Foundation.

References

1. Hilbert, M.: Big Data for development: a review of promises and challenges. *Dev. Policy Rev.* **34**(1), 135–174 (2015)
2. Wang, L., et al.: Bigdatabench: a big data benchmark suite from internet services. In: *Proceedings of 20th IEEE International Symposium on High Performance Computer Architecture*, pp. 488–499 (2014)

3. Li, D.R., Yao, Y., Shao, Z.F.: Big Data in the smart city. *Geomatics Inf. Sci. Wuhan Univ.* **39**(6), 630–640 (2014)
4. Al-Jarrah, O.Y., Yoo, P.D., Muhaidat, S., Karagiannidis, G.K., Taha, K.: Efficient machine learning for big data: a review. *Big Data Res.* **2**(3), 87–93 (2015)
5. Phatarfod, R.M.: Some aspects of stochastic reservoir theory. *J. Hydrol.* **30**(3), 199–217 (1976)
6. Bohling, G.: Stochastic simulation and reservoir modeling workflow. *Aust. J. Basic Appl. Sci.* **3**, 330–341 (2005)
7. Karacan, C.Ö., Olea, R.A.: Stochastic reservoir simulation for the modeling of uncertainty in coal seam degasification. *Fuel* **148**, 87–97 (2015)
8. Browning, C., Kumin, H.: Stochastic reservoir systems with different assumptions for storage losses. *Am. J. Oper. Res.* **6**(5), 414 (2016)
9. Archibald, T.W., McKinnon, K.I.M., Thomas, L.C.: An aggregate stochastic dynamic programming model of multi-reservoir systems. *Water Resour. Res.* **33**(2), 333–340 (1997)
10. Thomas, A., McMahan, T.A., Pegram, G.S., Vogel, R.M., Peel, M.C.: Revisiting reservoir storage-yield relationships using a global stream flow database. *Adv. Water Resour.* **30**, 1858–1872 (2007)
11. Prabhu, N.U.: Some exact results for the finite dam. *Ann. Math. Stat.* **29**(4), 1234–1243 (1958)

An Hybrid Multi-Core/GPU-Based Mimetic Algorithm for Big Association Rule Mining

Youcef Djenouri^{1,2(✉)}, Asma Belhadi¹, Philippe Fournier-Viger¹,
and Jerry Chun-Wei Lin³

¹ School of Humanities and Social Sciences, Harbin Institute of Technology,
Shenzhen, China

y.djenouri@gmail.com, abelhadi@usthb.dz, philfv8@yahoo.com

² Department of Computer Science,

Ulsan Institute of Technology, Ulsan, South Korea

³ School of Computer Science and Technology, Harbin Institute of Technology
Shenzhen Graduate School, Shenzhen, China

jerrylin@ieee.org

Abstract. This paper addresses the problem of big association rule mining using an evolutionary approach. The mimetic method has been successfully applied to small and medium size databases. However, when applied on larger databases, the performance of this method becomes an important issue and current algorithms have very long execution times. Modern CPU/GPU architectures are composed of many cores, which are massively threaded and provide a large amount of computing power, suitable for improving the performance of optimization techniques. The parallelization of such method on GPU architecture is thus promising to deal with very large datasets in real time. In this paper, an approach is proposed where the rule evaluation process is parallelized on GPU, while the generation of rules is performed on a multi-core CPU. Furthermore, an intelligent strategy is proposed to partition the search space of rules in several independent sub-spaces to allow multiple CPU cores to explore the search space efficiently and without performing redundant work. Experimental results reveal that the suggested approach outperforms the sequential version by up to at 600 times for large datasets. Moreover, it outperforms the-state-of-the-art high performance computing based approaches when dealing with the big WebDocs dataset.

Keywords: Mimetic algorithm · Association rule mining · Multi-core algorithm · GPU algorithm · Big data

1 Introduction

The goal of Association Rule Mining (ARM) is to discover all hidden patterns in a large transactional database [1]. The ARM problem can be formally described as follows. Let I be a set of n items $\{I_1, I_2, \dots, I_n\}$ and D be a set of m transactions $\{d_1, d_2, \dots, d_m\}$. An association rule ($X \Rightarrow Y$) is composed of two disjoint

parts, the set of items X , called the antecedent, and the set of items Y , called consequent. The ARM problem consists of finding all relevant association rules in a transactional database D . Two measures (support and confidence) are used in most ARM algorithms for evaluating association rules. They are based on the frequency of the items that appear in each given rule in a transaction database. The support of a rule is the number of transactions that contains its items over the number of all transactions. The confidence of a rule is the number of transactions that contains its items, divided by the number of transactions containing its antecedent. The goal of association rule mining is to discover strong rules, with both confidence and support beyond user's specific thresholds.

Many algorithms have been designed for ARM. Some algorithms are based on exact approaches like Apriori [1], FP-Growth [2]. Approaches based on the Apriori heuristic [1] first generate the k candidate itemsets from the $k - 1$ frequent itemsets and then test the frequency of the generated candidate itemsets, while approaches based on the FP-Growth heuristic [2] compress the transactional database in main memory using an efficient tree structure, and then recursively apply the mining procedure to find frequent itemsets. The exact approaches have long execution times for large transactional databases. Many other evolutionary-based algorithms have been developed to reduce the overall runtime mining process. Some of the main swarm intelligence approaches are, PeARM [3], IBSO-ARM [4], and some of the evolutionary approaches are GAR [5], and IARMMA [6]. The experimental study reported in [6] reveals that IARMMA outperforms the state-of-the-art ARM algorithms.

The existing association rule mining algorithms are high time consuming when dealing with massive data. Consequently, the ARM community has investigated the discovery of association rules in big data using different parallel hardware platforms, e.g. Cuda-APRIORI [7] and PBSO-ARM [8] for GPU computing, NGEF [9] and [11] for cluster computing and DARM for multi-core architecture [10]. However, these algorithms are still high time consuming for big data datasets. This is because parameters of a system's architecture (CPU/GPU communication for GPU computing, synchronization for Multi-core and cluster architecture) have a significant impact on mining performance.

To deal with big association rule mining, this paper proposes a hybrid multi-core/GPU based approach for mimetic ARM algorithm. The proposed approach relies on an efficient strategy to partition the search space of rules into smaller independent sub-space during the rule generation step using multiple CPU cores, and from the massively threaded CPU/GPU communication. To validate the proposed approach, several experiments have been carried out using big transactional databases. The results show that the proposed approach outperforms the state-of-the-art HPC-based ARM algorithms.

The reminder of the paper is organized as follows: Sect. 2 presents a brief explanation of the IARMMA algorithm. Section 3 presents the proposed PIARMMA algorithm. Section 4 evaluates the performance of the proposed algorithm using large and big datasets. Finally, Sect. 5 concludes the paper with some remarks and perspectives for future work.

2 IARMMA Algorithm

The IARMMA [6] algorithm (Improved Association Rule Mining using Mimetic Algorithm) was shown to outperform the state-of-the-art ARM algorithms. It first randomly generates an initial population of *pop_size* individuals. Then, the algorithm deletes each non admissible solution and applies a decomposition strategy on it. To keep the same population size, all individuals are evaluated using a fitness function. Then, IARMMA keeps only the best *pop_size* individuals (the others are removed). As the classical genetic algorithm, IARMMA applies the crossover and mutation operators. Afterwards, a local search is performed from each generated solution. The local search process is performed by applying successively a neighborhood computation process. Then, the delete and decomposition strategy is applied to eliminate non admissible rules.

Although IARMMA outperforms the state-of-the art ARM algorithms, it is high time consuming when dealing with large transactional databases. To deal with challenging issue, a parallel IARMMA algorithm is proposed in the next section.

3 Parallel IARMMA on GPU

The proposed parallel IARMMA algorithm first utilizes each CPU/core to generate a population. Then, the classical IARMMA algorithm is applied on each CPU/core. This includes performing the crossover, mutation, local search operations, and applying the delete and decomposition strategies, on multiple CPU/cores. For each iteration of the algorithm, many rules may be generated. For this reason the evaluation process is CPU high time consuming, especially when dealing with big data instances.

To deal with this issue, rules are evaluated using the GPU. In our previous study, experimental tests revealed that the evaluation of a single rule at a time is expensive on a GPU and requires CPU/GPU communication, which degrades the overall mining performance. To reduce CPU/GPU communication and to benefit from the massively threaded GPU, groups of multiple rules are submitted to the evaluation. Moreover, each block of threads is mapped to one rule. Threads of the same block are launched to calculate collaboratively the fitness of a single rule. Therefore, there are as many rules as blocks. The transactions are subdivided into subsets and each subset is assigned to exactly one thread so that each thread calculates only this part of the transactions set. After that a sum reduction is applied to aggregate the fitness value. At this stage, the GPU host sends all the fitness value of all rules of the CPU/cores. Afterwards, the CPU/cores select the best rules for the next iteration of the rule generation process. This mechanism is repeated until the maximum number of iterations is reached.

Using this approach a big search space of association rules may be explored. However, the CPU/cores may perform redundant work. Indeed, two or more CPU/cores can generate a same association rule. To deal with this issue, an efficient and fast heuristic is proposed to intelligently partition the big association

rule search space so that CPU/cores can work independently while avoiding performing redundant work.

Definition 1. *The status of an item for a given rule r must be one of the following three statuses:*

- *It appears in the antecedent of r .*
- *It appears in the consequent of r .*
- or*
- *it does not appear in the rule r .*

Based on Definition 1, the big rule space can be partitioned into several independent regions $R = \{R_1, R_2, \dots, R_{3 \times n}\}$, where each region R_i contains the set of rules that fix the value of the item $t_{\frac{i}{3}}$ to 0 (for indicating the absence of the item), 1 (for its presence in the antecedent) or 2 (for its presence in the consequent). Moreover, the pseudo code of the generation and the evaluation steps are given in Algorithms 1 and 2.

Algorithm 1. CPU/multi-core generation

```

1: PartitionBigRuleSpace( $R, R_1, R_2, \dots, R_p$ ).
2: for  $i = 0$  to IMAX do
3:   for  $j = 0$  to  $p$  cores do
4:     Crossover( $R_j, cr_i$ ).
5:     Mutation( $R_j, mr_i$ ).
6:     LocalSearch( $R_j, N_i$ ).
7:     DeleteDecompositionStrategy( $R_j$ ).
8:     cudaMemcpy( $R_j$ , cudaMemcpyHostToDevice).
9:   end for
10:  cudaMemcpy( $fitness(R_j)$ , cudaMemcpyDeviceToHost).
11:  selection( $R_j$ ).
12: end for

```

4 Performance and Results

This section reports on experiments, aiming to evaluate the performance of the proposed approach. The tests have been executed using the following large and big transactional databases:

- The databases are well-known scientific databases, frequently used in the data mining community (obtained from the FIMI dataset repository¹). The number of transactions for these datasets varies between 60,000 and 500,000 transactions, whereas, the number of items varies between 500 and 16,500 items (cf. Table 1).

¹ <http://fimi.ua.ac.be/data>.

Algorithm 2. GPU-based evaluation

```

1:  $idx \leftarrow \text{blockIdx.x} \times \text{blockDim.x} + \text{threadIdx.x}$ .
2: Compare the rule  $Buf f[\text{blockIdx.x}]$  and the transaction  $t_{idx}$ .
3: for  $i = 0$  to  $l$  transactions do
4:   if  $Buf f[\text{blockIdx.x}] \in t_{(i \times \text{blockDim.x}) + idx}$  then
5:      $\text{count}[\text{blockIdx.x}][i] \leftarrow 1$ .
6:   else
7:      $\text{count}[\text{blockIdx.x}][i] \leftarrow 0$ .
8:   end if
9: end for
10:  $\text{fitness}(Buf f[\text{blockIdx.x}]) \leftarrow \text{Sum\_Reduction}(\text{count}[\text{blockIdx.x}])$ .
11:  $\text{cudaMemcpy}(\text{fitness}(Buf f[\text{blockIdx.x}]), \text{cudaMemcpyDeviceToHost})$ .

```

- A big transactional database called Webdocs is also used. It contains more than 1,500,000 transactions and more than 500,000 items. To the best of our knowledge, the WebDocs dataset have never been used for discovering patterns using sequential ARM algorithms.

The proposed approach has been implemented using C-CUDA 4.0 and the experiments have been carried out using a CPU host coupled with a GPU device. The CPU host is a 64-bit quad-core Intel Xeon E5520 with a clock speed of 2.27 GHz with 8 cores. The GPU device is an Nvidia Tesla C2075 with 448 CUDA cores (14 multiprocessors with 32 cores each), a clock speed of 1.15 GHz, a 2.8 GB of global memory, 49.15 KB of shared memory, and a warp size of 32. Both the CPU and GPU are used in single precision.

Table 1. Description of the large datasets

Instance name	Transactions size	Item size	Average size
BMS-WebView-1	59,602	497	2.5
BMS-WebView-2	77,512	3,340	5
Korasak	80,769	7,116	50
Retail	88,162	16,469	10
Connect	100,000	999	10
BMP POS	515,597	1,657	2.5

In the following, several tests have been done to evaluate the proposed approach. First, its performance is compared with the sequential IARMMA in terms of execution times using large databases. Second, our approach is compared with the state-of-the-art HPC-based ARM approaches using the big WebDocs transactional database.

Figure 1 (1) presents the speed up of the proposed approach compared to the sequential version using large transactional databases. This figure shows that

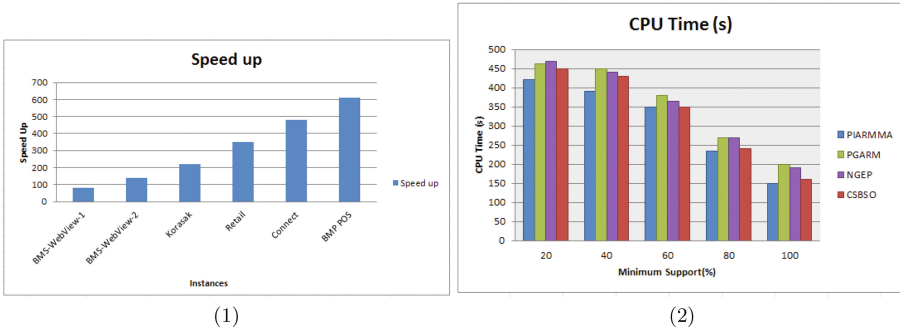


Fig. 1. Speed up of the proposed approach using large databases (1), The proposed approach Vs. the state of the art HPC-based ARM approaches in terms of runtime (s) using big dataset (WebDocs) with different minimum support (2).

the speed up increases as the number of transactions is increased. Indeed, for BMS-WEBVIEW-1, the speed up is 80 times (which means that the proposed approach is 80 times faster than the sequential version of IARMMA), whereas for BMS POS dataset, the speed up exceeds 600 times.

Figure 1 (2) presents the runtime of the proposed approach compared to the state of the art HPC-based ARM approaches, using the big WebDocs dataset. The results reveal that the proposed approach outperforms the state of the art HPC-based ARM approaches (CSBSO [11], PGARM [12], and NGEF [9]) for all minimum support threshold values used. Moreover, by increasing the minimum support from 20% to 100%, the runtime of HPC-based ARM approaches decreases.

Based on these results, it can be concluded that the proposed approach is competitive for big association rule mining. The approach benefits from the high computing power of CPU/multi-cores and GPU. Furthermore, results show that the strategy for partitioning the search space into independent sub-spaces is effective for distributing the workload between several CPU/cores to avoid that cores perform redundant work during the rule generation step.

5 Conclusion

In this paper, a new hybrid parallel algorithm based on the mimetic approach was presented for mining association rules in large and big databases. The pattern discovery process is distributed among several CPU cores, where each one explores independently a region of the search space thanks to a novel search space partitioning strategy. For big association rule mining, a GPU-based parallel algorithm is developed to evaluate the generated rules. The GPU evaluates multiple rules simultaneously such that each rule is mapped to one block of threads. To validate the usefulness of our approach, several experiments have been carried out on large and big datasets. Results show that the proposed approach outperforms the sequential version of the algorithm by up to 600 times on

large instances. The results also reveal that on the big Webdocs instance, the proposed approach outperforms the state-of-the-art HPC-based approaches. For future work, we plan to design other HPC based algorithms for big association rule mining in real-time.

References

1. Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large databases. *ACM SIGMOD Rec.* **22**(2), 207–216 (1993). ACM
2. Han, J., Pei, J., Yin, Y.: Mining frequent patterns without candidate generation. *ACM SIGMOD Rec.* **29**(2), 1–12 (2000). ACM
3. Gheraibia, Y., Moussaoui, A., Djenouri, Y., Kabir, S., Yin, P.Y.: Penguins search optimisation algorithm for association rules mining. *J. Comput. Inf. Technol. CIT* **24**(2), 165–179 (2016)
4. Djenouri, Y., Drias, H., Habbas, Z.: Bees swarm optimisation using multiple strategies for association rule mining. *Int. J. Bio-Inspired Comput.* **6**(4), 239–249 (2014)
5. Mata, J., Alvarez, J.L., Riquelme, J.C.: An evolutionary algorithm to discover numeric association rules. In: *Proceedings of the 2002 ACM Symposium on Applied Computing*, pp. 590–594. ACM (2002)
6. Djenouri, Y., Bendjoudi, A., Nouali-Taboudjemat, N., Habbas, Z.: An improved evolutionary approach for association rules mining. In: *Bio-Inspired Computing-Theories and Applications*, pp. 93–97. Springer, Heidelberg (2014)
7. Silvestri, C., Orlando, S.: GPUDCI: exploiting GPUS in frequent itemset mining. In: *2012 20th Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP)*, pp. 416–425. IEEE (2012)
8. Djenouri, Y., Bendjoudi, A., Habbas, Z., Mehdi, M., Djenouri, D.: Reducing thread divergence in GPU-based bees swarm optimization applied to association rule mining. *Concurrency Comput. Pract. Exp.* **29**(9) (2017)
9. Chen, Y., Li, F., Fan, J.: Mining association rules in big data with NGEF. *Cluster Comput.* **18**(2), 577–585 (2015)
10. Yoo, J.S., Boulware, D., Kimmey, D.: *Incremental and Parallel Association Mining for Evolving Spatial Data: A Less Iterative Approach on MapReduce* (2015)
11. Djenouri, Y., Bendjoudi, A., Djenouri, D., Habbas, Z.: Parallel BSO algorithm for association rules mining using master/worker paradigm. In: *International Conference on Parallel Processing and Applied Mathematics*, pp. 258–268. Springer International Publishing (2015)
12. Djenouri, Y., Bendjoudi, A., Djenouri, D., Comuzzi, M.: GPU-based bio-inspired model for solving association rules mining problem. In: *2017 25th Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP)*, pp. 262–269. IEEE (2017)

Updating the Discovered High Average-Utility Patterns with Transaction Insertion

Tsu-Yang Wu^{1,2}, Jerry Chun-Wei Lin^{3(✉)}, Yanan Shao³,
Philippe Fournier-Viger⁴, and Tzung-Pei Hong^{5,6}

¹ Fujian Provincial Key Laboratory of Big Data Mining and Applications,
Fujian University of Technology, Fuzhou 350118, China
wutsuyang@gmail.com

² National Demonstration Center for Experimental Electronic Information and
Electrical Technology Education, Fujian University of Technology, Fuzhou, China

³ School of Computer Science and Technology, Harbin Institute of Technology
Shenzhen Graduate School, Shenzhen, China
jerrylin@ieee.org, shaoyin0817@gmail.com

⁴ School of Natural Sciences and Humanities, Harbin Institute of Technology
Shenzhen Graduate School, Shenzhen, China
philfv@hit.edu.cn

⁵ Department of Computer Science and Information Engineering, National
University of Kaohsiung, Kaohsiung, Taiwan
tphong@nuk.edu.tw

⁶ Department of Computer Science and Engineering, National Sun Yat-sen
University, Kaohsiung, Taiwan

Abstract. In this paper, we propose an algorithm to handle the transaction insertion for efficiently updating the discovered high average-utility upper-bound itemsets (HAUUBIs) based on the average-utility (AU)-list structure and the Fast UPdated (FUP) concept. The proposed algorithm divides the HAUUBIs existing in the original database and new transactions into four cases, and each case can be respectively maintained to identify the actual high average-utility itemsets (HAUIs) without multiple database scans and enormous candidate generation. Experiments showed that the proposed algorithm has better performance compared to state-of-the-art algorithm in terms of runtime and generates the similar number of candidates.

Keywords: Average-utility list · HAUIM · Incremental · Dynamic database · Insertion

1 Introduction

Data mining is an emerging topic in recent decades since it can be used to discover the potential and implicit knowledge from a very large database, which can be used for decision making [1, 6]. Association-rule mining (ARM) or frequent-itemset mining (FIM) is the fundamental method to mine the relationship and

correlation between items. Agrawal et al. [1] first proposed Apriori algorithm to mine association rules in the level-wise manner with candidate generation approach. Han et al. [6] then compressed the databases into an frequent-pattern (FP)-tree structure and used FP-growth algorithm to mine the complete set of frequent patterns.

In ARM or FIM, a fundamental limitation is that the item only appears once in a transaction and all items have equal weight and importance. To address this issue, the high-utility itemset mining (HUIM) [8, 10, 16] was proposed to discover profitable itemsets from the database. Transaction-weighted utility (TWU) model [9] was first presented to hold the downward closure property for mining the high-utility itemsets (HUIs). Several algorithms were extensively studied and still developed in progress [2, 4, 11, 14]. In HUIM, the utility of an itemset increases along with the length of it (number of items within the itemset). Thus, longer itemset tends to obtain higher profit, which is an unfair measure to identify the high-utility itemsets. To address this issue, Hong et al. [7] proposed high average-utility itemset mining (HAUIM) to take the length of the itemset into account for measuring the average-utility of the itemset, thus providing a fair assessment to the itemset. Several algorithms were presented but most of them focus on mining high average-utility itemsets (HAUIs) in the static database [12, 13, 15].

In real-life situations, the database usually changes all the time. In this paper, we thus present an efficient algorithm to efficiently update the discovered HAUIs based on the average-utility (AU)-list structure [12] and Fast Updated (FUP) concept [3] with transaction insertion. The proposed algorithm can efficiently update the discovered information to identify the actual HAUIs in the dynamic situation.

2 Preliminaries and Problem Statement

A quantitative database is shown in Table 1. Each item contains the purchase quantity of it in each transaction. An example is then shown in Table 1. The profit table indicates the unit profit of each item appearing in the database, which can be defined as $\{a:4, b:6, c:5, d:9, e:12, f:8, g:3\}$ for the running example. A minimum high average-utility threshold δ is set according to the user's preference (a positive integer). For example, the minimum high average-utility threshold in this example is set as 16%.

Definition 1. The average-utility of an item i_j in a transaction T_q is denoted as $au(i_j, T_q)$, and defined as:

$$au(i_j, T_q) = \frac{q(i_j, T_q) \times p(i_j)}{1}, \quad (1)$$

where $q(i_j, T_q)$ is the quantity of i_j in T_q , and $p(i_j)$ is the unit profit value of i_j .

Table 1. A quantitative database.

TID	Items with their quantities
T_1	$a:5, b:2, c:3, d:2$
T_2	$b:5, d:4, e:2, g:12$
T_3	$a:1, c:5, f:4$
T_4	$c:2, d:3, e:3$
T_5	$a:3, d:2, f:6$

Definition 2. The average-utility of a k -itemset X in a transaction T_q is denoted as $au(X, T_q)$, and defined as:

$$au(X, T_q) = \frac{\sum_{i_j \subseteq X \wedge X \subseteq T_q} q(i_j, T_q) \times p(i_j)}{|X|} = \frac{\sum_{i_j \subseteq X \wedge X \subseteq T_q} q(i_j, T_q) \times p(i_j)}{k}, \quad (2)$$

where k is the number of items in X .

Definition 3. The average-utility of an itemset X in D is denoted as $au(X)$, and defined as:

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

Definition 4. The transaction utility of a transaction T_q is denoted as $tu(T_q)$, and defined as:

$$tu(T_q) = \sum_{i_j \subseteq T_q} u(i_j, T_q). \quad (4)$$

Definition 5. The total utility of a database D is denoted as TU , and defined as the sum of all transaction utilities, that is:

$$TU = \sum_{T_q \in D} tu(T_q). \quad (5)$$

Definition 6 (Transaction-maximum utility). The transaction-maximum utility of a transaction T_q denoted as $tmu(T_q)$ and defined as:

$$tmu(T_q) = \max\{u(i_j) | i_j \subseteq X \wedge X \subseteq T_q\}. \quad (6)$$

Definition 7 (Average-utility upper-bound, $auub$ property). The average utility upper bound of an itemset X is denoted as $auub(X)$ and defined as:

$$auub(X) = \sum_{X \subseteq T_q \wedge T_q \in D} tmu(T_q), \quad (7)$$

where $tmu(T_q)$ is the maximum utility of transaction T_q such that $i_j \subseteq X \wedge X \subseteq T_q$.

Definition 8 (High average-utility upper bound itemset, *HAUUBI*).

An itemset X is a HAUUBI if it satisfies the condition as:

$$HAUUBI \leftarrow \{X | auub(X) \geq TU^D \times \delta\} \quad (8)$$

Assume that the newly inserted database shown in Table 2 is used as the inserted transactions. The problem definition can be defined as follows.

Problem Statement: The problem of incremental HAUM with transaction insertion is to efficiently update the set of the HAUUBIs to identify the actual HAUIs. Thus, the updated HAUI can be formally defined as:

$$HAUI \leftarrow \{X | au(X)^U \geq (TU^D + TU^d) \times \delta\}, \quad (9)$$

where $au(X)^U$ is the actual average-utility of X in the updated database, TU^D is the total utility in the original database, TU^d is the total utility in the inserted database (d), and δ is the minimum high average-utility threshold, which can be specified by user's preference.

Table 2. Newly inserted transactions.

TID	Items
T_6	$a:4, b:6, c:4, d:2$
T_7	$a:5, c:3, e:1, f:2, g:6$

3 Proposed Incremental Algorithm

The designed algorithm mainly has two phases. In the first phase, the set of 1-HAUUBIs are discovered from the original database. The set of 1-HAUUBIs in the original database was the high average-utility upper-bound 1-itemsets, which can be used to maintain the downward property of HAUM. The set of 1-HAUUBIs in the original database was then used to build the AU-list structure [12] for later processing. In the second phase, when some transactions are inserted into the original database, the designed algorithm then divides the sets of HAUUBIs found in original database and new transactions into four parts based on the Fast UPdated (FUP) concept [3]. The four cases of the designed approach is shown in Fig. 1.

The set of HAUUBIs of each part is then maintained by the designed procedures except the itemsets in Case 4. The reason is that the itemsets in Case 4 cannot be the HAUIs while the database is updated. Details of the designed algorithm is shown in Algorithm 1.

The **updateADD** and **updateDEL** approaches are used to respectively add and delete the items in the AU-list structure. The **updateADD** function can easily update the *auub* value of the itemsets based on the AU-list structure. For the **updateDEL** function, it can directly remove the unpromising itemsets based on the AU-list structure after database is updated. After that, the

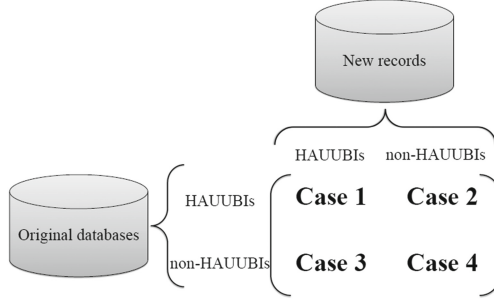


Fig. 1. Four cases of the designed algorithm with transaction insertion.

Algorithm 1. Proposed algorithm

Input: D , a quantitative database; $ptable$, a profit table; the total utility TU^D in D ; δ , the minimum high average-utility threshold; AUL , the built AU-list from D ; d , a set of inserted transactions.

Output: the sets of HAUUs and HAUUBIs.

```

1 set  $HAUUBIs.U \leftarrow null$ ;
2 calculate  $TU^d$  in  $d$ ;
3 for each  $i_j \in d$  do
4   calculate  $auub(i_j)$ ;
5   if  $auub(i_j)^d \geq TU^d \times \delta$  then
6      $1-HAUUBIs.d := 1-HAUUBIs.d \cup i_j$ ;
7  $TU^U := TU^D + TU^d$ ;
8 for each  $i_j \in T_q \subseteq d$  do
9   if  $i_j \in 1-HAUUBIs.D$  then
10     $auub(i_j)^U := auub(i_j)^D + auub(i_j)^d$ ;
11    if  $auub(i_j)^U \geq (TU^D + TU^d) \times \delta$  then
12      updateADD( $AUL$ );
13       $HAUUBIs.U := HAUUBIs.U \cup i_j$ ;
14    else
15      updateDEL( $AUL$ );
16  else
17     $scan\_set := scan\_set \cup i_j$ ;
18 for  $i_j \in scan\_set$  do
19   calculate  $auub(i_j)^D$ ;
20   calculate  $auub(i_j)^U := auub(i_j)^D + auub(i_j)^d$ ;
21   if  $auub(i_j)^U \geq TU^U \times \delta$  then
22     updateADD( $AUL$ );
23   if  $AUL \neq null$  then
24     Construct( $AUL$ );
25     update  $HAUUBIs.U$ ;
26   identify the set of  $HAUUs$  from  $HAUUBIs.U$ ;

```

remaining itemsets in the AU-list are then checked against to the minimum high average-utility count in the updated database, and the actual high average-utility itemsets can thus be maintained.

4 Experimental Results

In this section, the performance of the proposed algorithm is compared with the stat-of-the-art HAU-Miner [12]. All algorithms were implemented in Java and performed on a personal computer with an Intel(R) Core (TM) i7-6700 4.00 GHz processor and 8 GB of the memory, running on the 64 bit Microsoft Windows 10 operating system. Experiments were conducted on one real-world and one synthetic datasets [5]. To assess the performance of the proposed algorithm, the runtime and the number of candidates of two algorithms are then compared and evaluated.

4.1 Runtime

In this experiments, the runtimes of two algorithms are then compared under varied minimum high average-utility thresholds with a fixed insertion ratio. The results are shown in Fig. 2.

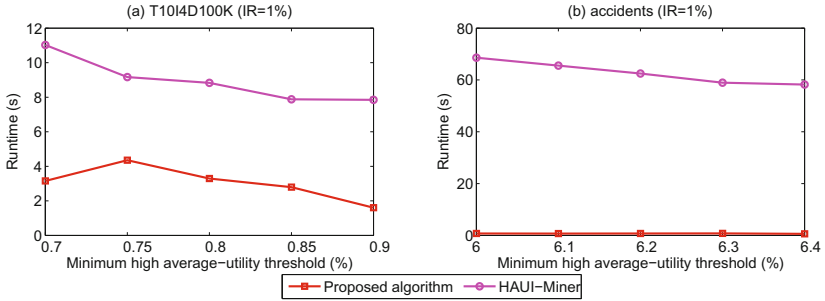


Fig. 2. Runtimes under varied minimum high average-utility thresholds.

From Fig. 2, it can be seen that the runtime decreases along with the increasing of the minimum high average-utility threshold. This is reasonable since when the minimum high average-utility threshold increases, less computational time is required to mine fewer HUIs. Also, it can be observed that the designed algorithm outperforms the state-of-the-art HAU-Miner approach for updating the HAUUBIs to identify the actual HAUIs.

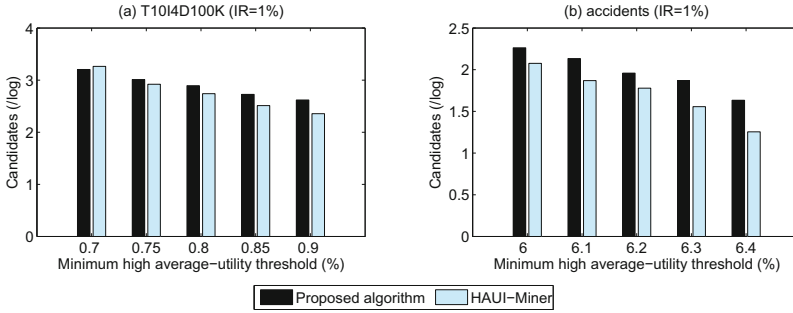


Fig. 3. Numbers of candidates under varied minimum high average-utility thresholds.

4.2 Number of Candidates

In this section, we compare two algorithms in terms of number of candidates under two datasets. Results are shown in Fig. 3.

From Fig. 3, we can see that the numbers of candidates of two algorithms are nearly similar since those two approaches are based on the AU-list structure. The designed algorithm sometimes requires more candidates for evaluation. This is reasonable since more HAUUBIs were kept for later updating. However, with little more candidates, the runtime of the designed algorithm greatly outperforms the state-of-the-art approach.

5 Conclusion

In this paper, we present an algorithm to update the discovered HAUUBIs then identify the actual HAUIs based on the efficient average-utility (AU)-list structure. This structure can be easily to maintain and reduce the cost of multiple database scans without generating the enormous candidates. From the experimental evaluation, it can be obvious to see that the proposed algorithm outperforms the previous state-of-the-art approach.

Acknowledgment. This research was partially supported by the National Natural Science Foundation of China (NSFC) under grant No. 61503092, by the Shenzhen Technical Project under JCYJ20170307151733005, by the Science Research Project of Guangdong Province under grant No. 2017A020220011, and by the National Science Funding of Guangdong Province under Grant No. 2016A030313659.

References

1. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: International Conference on Very Large Data Bases, pp. 487–499 (1994)

2. Ahmed, C.F., Tanbeer, S.K., Jeong, B.S., Lee, Y.K.: Efficient tree structures for high utility pattern mining in incremental databases. *IEEE Trans. Knowl. Data Eng.* **21**(12), 1708–1721 (2009)
3. Cheung, D.W., Wong, C.Y., Han, J., Ng, V.T.: Maintenance of discovered association rules in large databases: an incremental updating techniques. In: *The International Conference on Data Engineering*, pp. 106–114 (1996)
4. Erwin, A., Gopalan, R. P., Achuthan, N. R.: Efficient mining of high utility itemsets from large datasets. In: *The Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, pp. 554–561 (2008)
5. Fournier-Viger, P., Lin, J.C.W., Gomariz, A., Gueniche, T., Soltani, A., Deng, Z., Lam, H.T.: The SPMF open-source data mining library version 2 and beyond. In: *The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery*, pp. 36–40 (2016)
6. Han, J., Jian, P., Yin, Y., Mao, R.: Mining frequent patterns without candidate generation: a frequent-pattern tree approach. *Data Mining Knowl. Discov.* **8**(1), 53–87 (2004)
7. Hong, T.P., Lee, C.H., Wang, S.L.: Effective utility mining with the measure of average utility. *Expert Syst. Appl.* **38**(7), 8259–8265 (2011)
8. Liu, Y., Liao, W.K., Choudhary, A.: A fast high utility itemsets mining algorithm. In: *International Workshop on Utility-Based Data Mining*, pp. 90–99 (2005)
9. Liu, Y., Liao, W.K., Choudhary, A.: A two-phase algorithm for fast discovery of high utility itemsets. In: *The Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, pp. 689–695 (2005)
10. Lin, C.W., Hong, T.P., Lu, W.H.: An effective tree structure for mining high utility itemsets. *Expert Syst. Appl.* **38**(6), 7419–7424 (2011)
11. Liu, M., Qu, J.: Mining high utility itemsets without candidate generation. In: *ACM International Conference on Information and Knowledge Management*, pp. 55–64 (2012)
12. Lin, J.C.W., Li, T., Fournier-Viger, P., Hong, T.P., Zhan, J., Voznak, M.: An efficient algorithm to mine high average-utility itemsets. *Adv. Eng. Inform.* **30**(2), 233–243 (2016)
13. Lin, J.C.W., Ren, S., Fournier-Viger, P., Hong, T.P., Su, J.H., Vo, B.: A fast algorithm for mining high average-utility itemsets. *Appl. Intell.* **41**(2), 331–346 (2017)
14. Lin, J.C.W., Gan, W., Fournier-Viger, P., Chao, H.C.: FDHUP: fast algorithm for mining discriminative high utility patterns. *Knowl. Inf. Syst.* **51**(3), 873–909 (2017)
15. Lin, J.C.W., Ren, S., Fournier-Viger, P., Hong, T.P.: EHAUPM: efficient high average-utility pattern mining with tighter upper-bounds. *IEEE Access* **5**, 12927–12940 (2017)
16. 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. 215–221 (2004)

Image and Multimedia Processing

Applying Image Processing Technology to Region Area Estimation

Yi-Nung Chung¹, Yun-Jhong Hu¹, Xian-Zhi Tsai¹,
Chao-Hsing Hsu^{2(✉)}, and Chien-Wen Lai³

¹ Department of Electrical Engineering, National Changhua
University of Education, Changhua 500, Taiwan
ynchung@cc.ncue.edu.tw

² Department of Information and Network Communications,
Chienkuo Technology University, Changhua 500, Taiwan
hsu@cc.ctu.edu.tw

³ Changhua Christian Hospital, Changhua 500, Taiwan
70672@cch.org.tw

Abstract. This paper proposes a method to measure a region area of field by using aerial images. An unmanned aerial vehicle (UAV) and image processing technology is used to capture images of the land and measure its area. The main advantage of using UAV to capture images is the higher degree of freedom; it can accord user's operation to capture from various angles and heights to obtain more diversified information. Even taking pictures of a dangerous area, the user can remote the UAV in a safer place, and get the information of the area or the UAV in real time. In the experiment, an UAV is used to get images of the playground grassland which region area is known, and capture a group of images with same area from 70 to 120 m height every ten meters. In image processing process, edge detection and morphology are used to find the range of the interest region, and then count the number of pixels of it. We can get the relation between the different height and per pixels of the real area. Experimental results show that the average deviations of estimating unknown area are less than 2%.

Keywords: Image processing · Unmanned aerial vehicle · Edge detection · Area estimation

1 Introduction

Because the measurement technology made progress, the method of measuring the land area is faster and easier. The method greatly improves the efficiency of the measurement area. General methods of measuring land area are traditional measurement, aerial photogrammetry and global positioning system measurement. The traditional measurement methods are the most convenient, but that are limited by weather and terrain effects. If we use aerial photogrammetry or global positioning system measurement, that can save manpower and time. Remotely Piloted Vehicle (RPV) [7, 8] has been actively developed in recent years. Some advanced countries have applied them to military aspects, such as enemy position detection and coastal investigation monitoring, etc. There are many

applications in the non-government institution too, such as the exploration of urban development, road traffic analysis and disaster area relief investigation. The main advantage of using unmanned aerial vehicle (UAV) to capture images is less susceptible to terrain. It can accord user's operation to capture from various angles and heights to obtain more diversified information. There is a lower cost and can be popularized by captured images. Compared with the traditional methods, the user can remotely control unmanned aerial vehicle in a safe place to capture image of the dangerous area, and get the information of the area in real time.

The land area is often affected by the land boundary, which makes the actually area of crop area different. In the estimation of rice planting area, growth area and production capacity will be affected. If the methods can calculate the actual area of planting area. It will be able to more accurately estimate the production of each crop. The mangrove area has been monitored and estimated to the land area by unmanned aerial vehicle. But the height of the unmanned aerial vehicle can only be estimated at a fixed height. This paper proposes a method to measure a region area of field by using aerial images. The method uses the UAV to collect the image of the grass ground area in each height interval and use the image processing to calculate the connected region pixels in the region of interest. That will obtain the relation of different height and per pixels of the real area. According to this data can be derived least squares regression line, so that we can estimate the unknown region area of field if we know the height of the aerial image. The research of this paper is divided into three parts: aerial image collection, the relation between the different height and per pixels of the real area, and finally the region area estimates.

This research proposes to use remote control unmanned aerial vehicle to capture multiple sets of aerial images, and collecting images ranging from 70 to 120 m height. Every 10 m collects 30 to 50 groups of the same height of the same area of the regional image. After collecting enough aerial images, it can process these images. In order to reduce the operation time, the original image is reduced to 25%. Using the lens distortion normalization and then select the region of interest to be processed. Through the color component conversion of the original input RGB color image is converted to gray Intensity image, histogram equalization and edge detection. After the number of pixels in the acquisition area, we can get the relation between the different height and per pixels of the real area. And the distribution of the relationship is obtained by the height. In order to estimate the distribution of each set of relational data, we propose to use the least square method to find the regression line. Therefore, we can use this equation to estimate the region effectively.

2 Image Processing

General camera to shoot the image for the RGB color image, if it directly processes of this image, it will spend a lot of computing time. So the system will convert RGB color image to grayscale image and then do following process. In general, the surveillance system shows the images by RGB color system but is easy influenced by shadow or ambient illumination changes. In this paper, we transform the RGB system to HSV color system which can reduce these influence. HSV color space [1, 2] contains three

components, in which H represents the hue, S represents the saturation, and V represents the brightness. The hue element usually thinks of as color like red, green or others. The saturation is the ratio of colorfulness to brightness which is just gray scale. The value is another word for color lightness like the light green or dark green. After the color space conversion algorithm, the system obtains a relative image with reference background. The image contains RGB or HSV three components, which will increase the computation burden. Therefore, it will do binarization process and let the picture become a gray image. Then the image will reduce to be gray image.

After the image is become grayscale, the environmental factors will lead to different image quality, especially in the outdoor image. The image enhancement is used to solve the above problem, which involves histogram equalization and edge sharpening [3, 4] processing. The results obtained by equalizing the histogram, it can be observed that the histogram distribution is more evenly distributed between 0 and 255 grayscale intensity. The image is more obvious than the original gray scale image and the contrast and brightness are significantly improved. The image features and details are more obvious, suitable for following process.

After the histogram is equalized, the contrast of the image is obviously increased, the features and details are also obvious. The results of the following edge detection need more accurate, so the image needs further edge enhancement program. In this paper, we use the Un-sharp Masking method in edge sharpening to enhance the edge. The concept is to subtract a fuzzy image from the original image to adjust the numerical scale so that the output image is more clear, and to achieve the edge of the enhanced effect. It can improve the accuracy of following process.

The purpose of edge detection is to detect the point where the brightness change in the image is obvious. It is usually the boundary of the object and the background. The gray scale image converted by the original image contains many irrelevant background noise. The edge detection method can be used to detect. If it is possible to detect the location of the block. Common edge detection methods include Sobel edge detection, Prewitt edge detection, and Canny edge detection method. The principle of the image based on the size of the gray scale difference and gradient degree of change to determine the greater the gap on behalf of its neighboring pixels. There are more obvious bright and dark changes. These positions are usually the object and the background of the border, so you can use this principle to detect the image of the edge of the object. In this paper, Canny edge detection method is used to detect the edge of the block to be detected. The steps are to filter out the noise by using a $3 * 3$ Gaussian filter, and then calculate the lateral and longitudinal differential approximation for each pixel to obtain the pixel. According to this method, the edge component in the image can be detected.

Morphology [5, 6] processing is often used in target detection, noise removal, block segmentation and skeleton boundary capture, etc. The principle is based on the mathematical theory of the collection. The operation is to use mask in the image of the pixels as a shift operation. This mask also known as structural elements, the user can set the size and shape of the structural elements. According to different Morphological algorithms do different treatments to achieve image segmentation and recognition purposes. The basic operations are Erosion, Dilation, Opening and Closing. Many of the applications of morphology can be deduced based on these basic operations to

perform advanced image processing. In addition, applying morphology to image processing simplifies image data and maintains the basic outline of the graph.

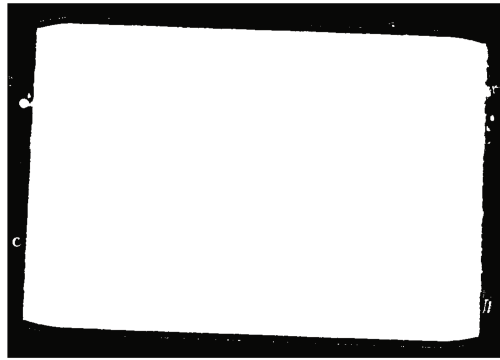
3 Experimental Results

The images used in this paper are taken by an unmanned aerial camera [7, 8], each of them is taken in the same area and contains a piece of artificial grassland area, as shown in Fig. 1. The center of the image is the center of the playground of the block and the artificial measurement of the block area for the reason. It is difficult to find the area of the existing region. The least squares regression equation can be used to obtain the area of artificial measurement area. Because the field of artificial measurement area of the playground grass is a rectangular block, so we can use of range finder for measurement. The height of the recorded images is between 70 and 120 m. The number of shots is about 30 to 50 times. The aim is to increase the accuracy of the regression line and then to analyze the images. After the image processing, we can get the binarized image of the measurement area. As shown in Fig. 2(a), the white area is the range of the measurement area and it can be observed that there are many minor noise next to it. In order to enhance the accuracy of the estimated area, so we use connectivity area marking method, which you can effectively extract the required number of block pixels. The calculated number of pixels will be displayed in the operating window shown in Fig. 2(b). By using this method, it can find in the largest block of the number of pixels, which is the playground grass.



Fig. 1. Artificial measurement area

After calculating the number of pixels in each image, we can get the relationship between the height and the number of pixels representing the actual area, as shown in Fig. 3, where the horizontal axis is the height and the vertical axis is the ratio of actual area and the pixel number. Since the distribution of each set of relational data is not



(a)

Variables - pixelnum

pixelnum x

pixelnum <1x79 double>

	1	2	3	4	5	6	7	8	9	10
1	44	94	38	9	12	30	37	582	2646891	9

(b)

Fig. 2. (a) Measuring area range (b) the number of pixels

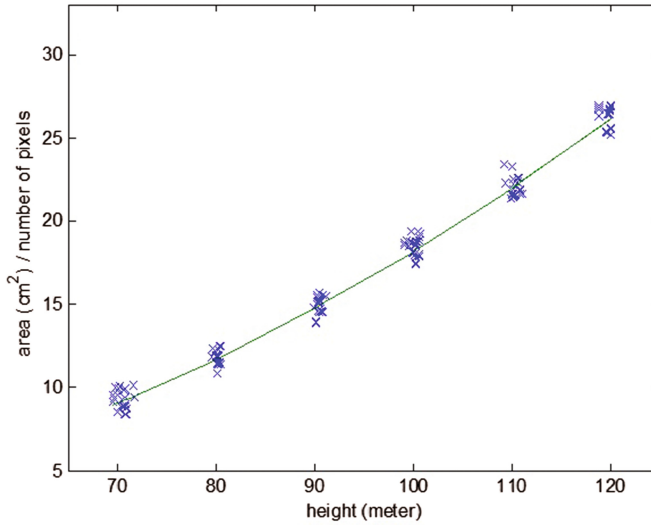


Fig. 3. Relationship between the number of pixels and the area of the distribution map and its regression line

Table 1. Estimation of field area

Real area	Estimated area	Error
260.87 m ²	265.5 m ²	1.77%

linear, in order to estimate the area of other different sizes, this study proposes to use the least square method [9]. Using the least square method, the error of the estimation can be reduced.

In the experiment, the regression equation was used to estimate the area of the specific area at different heights. The specific area estimation result by using the proposed method is shown in Table 1.

4 Conclusion

In this paper, an image processing technology is proposed to estimate the land area by using UAV images. Firstly, the image information is collected by the unmanned aerial vehicle, and the area of the region is to be detected by image processing methods which include edge detection, morphology, and the least squares regression line. The regression line can be used to obtain other unknown area. Compared with other methods of measuring area, this method has the advantages of simple system design and low cost. The experimental result shows that the error rate of the regression line is less than 2% on average.

Acknowledgments. This work was supported by the Ministry of Science and Technology under Grant MOST 103-2221-E-018-017- and MOST 105-2221-E-018-023-.

References

1. Rahman, M.A., Purnama, I.K.E., Purnomo, M.H.: Simple method of human skin detection using HSV and YCbCr color spaces. In: 2014 IEEE International Conference on Intelligent Autonomous Agents, Networks and Systems (INAGENTSYS), pp. 58–61 (2014)
2. Liu, F., Liu, X., Chen, Y.: An efficient detection method for rare colored capsule based on RGB and HSV color space. In: 2014 IEEE International Conference on Granular Computing (GrC), pp. 175–178 (2014)
3. Wang, W., He, Y., Li, Z., Chen, Z.: A real-time target detection algorithm for Infrared Search and track system based on ROI extraction. In: 2012 IEEE International Conference on Signal Processing, Communication and Computing (ICSPCC), pp. 774–778 (2012)
4. Yitzhaky, Y., Peli, E.: A method for objective edge detection evaluation and detector parameter selection. *IEEE Trans. Pattern Anal. Mach. Intell.* **25**(8), 1027–1033 (2003)
5. Qiu, T., Yan, Y., Gang, L.: An auto-adaptive edge-detection algorithm for flame and fire image processing. *IEEE Trans. Instrum. Meas.* **61**(5), 1486–1493 (2012)
6. Peng, B., Zhang, L., Zhang, D.: Automatic image segmentation by dynamic region merging. *IEEE Trans. Image Process.* **20**(12), 3592–3605 (2011)

7. Vladimir, T., Jeon, D., Kim, D.-H., Chang, C.-H., Kim, J.: Experimental feasibility analysis of ROI-based hough transform for real-time line tracking in auto-landing of UAV. In: 2012 15th IEEE International Symposium on Object/Component/Service-Oriented Real-Time Distributed Computing Workshops (ISORCW), pp. 130–135 (2012)
8. Bi, J., Mao, W., Gong, Y.: Research on image mosaic method of UAV image of earthquake emergency. In: Third International Conference on Agro-geoinformatics, Agro-geoinformatics 2014, pp. 1–6 (2014)
9. Larson, R., Falvo, D.C.: Elementary Linear Algebra, 6th edn. Brook/Cole Cengage Learning, Boston (2010)

Face Recognition under Lighting Variation Conditions Using Tan-Triggs Method and Local Intensity Area Descriptor

Chi-Kien Tran¹(✉), Duc-Tinh Pham¹, Chin-Dar Tseng²,
and Tsair-Fwu Lee^{2,3,4}(✉)

¹ Center for Information Technology, Hanoi University of Industry,
Hanoi, Vietnam

kientc.hau@gmail.com

² Medical Physics and Informatics Laboratory of Electronics Engineering,
National Kaohsiung University of Applied Sciences, Kaohsiung, Taiwan, ROC
tflee@kuas.edu.tw

³ Institute of Clinical Medicine, Kaohsiung Medical University,
Kaohsiung 807, Taiwan, ROC

⁴ Department of Radiation Oncology, Kaohsiung Chang Gung
Memorial Hospital, Kaohsiung, Taiwan, ROC

Abstract. Lighting variation is a specific and difficult case of face recognition. A good combination of an illumination preprocessing method and a local descriptor, face recognition system can considerably improve prediction performance. Recently, a new descriptor, named local intensity area descriptor (LIAD), has been introduced for face recognition in ideal and noise conditions. It has been proven to be insensitive to ideal and noise images and has low histogram dimensionality. However, it is not robust against illumination changes. To overcome this problem, in this paper, we propose an approach using an illumination normalization method developed by authors Tan and Triggs to normalize face images before encoding the processed images based on LIAD. The recognition was performed by a nearest-neighbor classifier with chi-square statistic as the dissimilarity measurement. Experimental results, conducted on FERET database, confirmed that our proposed approach performs better than traditional LIAD method and local binary patterns, local directional pattern, local phase quantization, and local ternary patterns using the same approach with respect to illumination variation.

Keywords: Face recognition · Illumination preprocessing · Tan-Triggs method · Local intensity area descriptor

1 Introduction

Varying lighting condition represents a significant technical challenge in face recognition technology [1–3]. It affects characteristics of face images. The difficulty of this problem stems from the fact that the captured images of a person under different lighting conditions are quite different because light can produce bright, dark, or shadowy regions

on facial images. A suitable combination of an illumination preprocessing method and a local descriptor will significantly improve the recognition rate of the face recognition under illumination variation conditions.

Numerous illumination preprocessing methods for addressing the illumination variation problems have been introduced [4–6]. In general, they can be separated into three major types: gray-level transformation (histogram equalization [7], logarithmic transform [8], and gamma intensity correction [9]), gradient and edge extraction (difference of Gaussians [10, 11] and gradient faces [6]), and face reflection field estimation (self-quotient image [12], local normalization [13], and Weber-face [5]). The output of these methods is still an image, so it needs be followed by a feature extractor method to extract the features of normalized images for classification purpose. In practical applications, none of existing illumination normalization methods can adequately deal with the effects of light on facial images and there is also no universally proper method for all local descriptors. If an illumination suppression method and a local descriptor can be appropriately combined, it is possible to improve the performance of a face recognition system.

Among several local descriptors that are insensitive to lighting variation, local binary pattern (LBP) [14, 15] is one of the most widely used in various computer vision problems and it has been proven to be a very efficient descriptor in face recognition under monotonic gray-scale changes [16]. However, it is not robust against local changes in the texture, caused by noise and illumination directions because it produces labels based on the relative intensity values of the center pixel and the pixels in the neighborhood [16, 17].

In order to overcome the drawback of LBP, Tan and Triggs proposed a three-level operator called local ternary patterns (LTP) to solve the problems of near constant image areas [18]. LTP labels a given pixel by three values (1, 0 or -1) based on the difference between the center pixel and a neighboring pixel according to a threshold T . It has provided better performance compared to LBP in face recognition and has been proven effective in face recognition under the monotonic illumination changes. However, it has a high histogram dimensionality, is sensitive to noise and illumination directions and the produced results are dependent on the value of threshold T .

In [19], authors proposed a descriptor based on quantizing the Fourier transform phase in local neighborhoods called local phase quantization (LPQ). It is known to be an illumination and blur insensitive feature extractor and obtain the better accuracy than LBP in face recognition [20]. However, it has still limitations about noise and strong lighting variation.

In [21], Jabid et al. introduced a descriptor based on edge extraction called local directional pattern (LDP). The authors have also indicated that the LDP descriptor is robust against varying lighting condition and aging effects compared to the LBP descriptor. However, the features used are only derivative, describing general trends, and the method has several drawbacks including high computational costs, noise sensitivity, variations in facial expression, and rotation of pose [17].

Recently, Tran et al. [17] proposed a new descriptor, named local intensity area descriptor (LIAD) for face recognition in ideal and noise conditions. It labels each pixel of a given image by a value that is the area of its eight-neighbor intensity values. It has been proven to be efficient to changes in ideal, noise and has low histogram

dimensionality compared with conventional descriptors such as LBP, LDP, LTP, and histogram of oriented gradients (HOG) [22], but be sensitive to variations in expression, aging, and illumination.

To reduce the effect of lighting variations in facial images, an illumination normalization method, developed by Tan and Triggs (Tan-Triggs or TT) [18], was used to normalize the light on face images before encoding the obtained images by LIAD. Experimental results, which were tested on a FERET [23] database, confirmed that this approach performs better than traditional LIAD method and LPQ, LDP, LPQ, and LTP using the same experimental conditions with respect to illumination variation.

The remaining part of this paper comprises of four sections. In Sect. 2, we introduce the LIAD and TT methods in detail. Section 3 presents face recognition using TT method and local intensity area descriptor. In Sect. 4, experiment settings, results, and some discussions are indicated. Finally, conclusions are drawn in Sect. 5.

2 Related Works

2.1 Local Intensity Area Descriptor

Tran et al. [17] proposed a local texture descriptor, named local intensity area descriptor (LIAD). LIAD labels a pixel of a given image by the area of its eight-neighbor intensity values and the obtained values are rounded down or up to an integer. The area is calculated based on a trapezoidal numerical integration formula [24], which is defined as

$$J = \int_a^b f(x)dx \approx \frac{h}{2}[f(a) + 2f(x_1) + 2f(x_2) + \dots + 2f(x_{n-1}) + f(b)], \quad (1)$$

where n is the number of rectangles and $h = (b - a)/n$. The x_j , a , and b are called nodes. In [17], the intensity value of pixels are considered as nodes and the value of h parameter is set by user.

Formally, the basic LIAD descriptor can be represented as

$$LIAD(x_c, y_c) = \frac{h}{2}(g_0 + 2g_1 + 2g_2 + \dots + 2g_6 + g_7), \quad (2)$$

where (x_c, y_c) is the central pixel of a 3×3 neighborhood and g_i ($i = 0, \dots, 7$) corresponds to the gray values of eight neighbors.

The basic LIAD is divided into two cases: $LIAD_{down}$ returns an integer after rounding down the result to the nearest integer towards minus infinity (Formula (3)); $LIAD_{up}$ returns an integer after rounding up the result to the nearest integer towards plus infinity (Formula (4)).

$$LIAD_{down}(x_c, y_c) = \text{floor}\left(\frac{h}{2}(g_0 + 2g_1 + 2g_2 + \dots + 2g_6 + g_7)\right), \quad (3)$$

$$LIAD_{up}(x_c, y_c) = \text{ceil}\left(\frac{h}{2}(g_0 + 2g_1 + 2g_2 + \dots + 2g_6 + g_7)\right), \quad (4)$$

where floor and ceil functions return an integer after rounding down and up the result, respectively.

An example of LIAD is illustrated by Fig. 1. Table 1 displays the maximal dimensionality of the LIAD histogram. The authors supposed that the gray-values of eight neighbors are the same value, 255. In this table, the predefined values of h parameter are shown in the first column; the corresponding maximum values of histogram dimensionality are shown in the second column; the maximal dimensionalities of the LIAD histogram (the nearest integers less than corresponding values in the second column) are displayed in the third column; the maximal dimensionalities of the LIAD histogram (the nearest integers greater than corresponding values in the second column) are displayed in the last column.

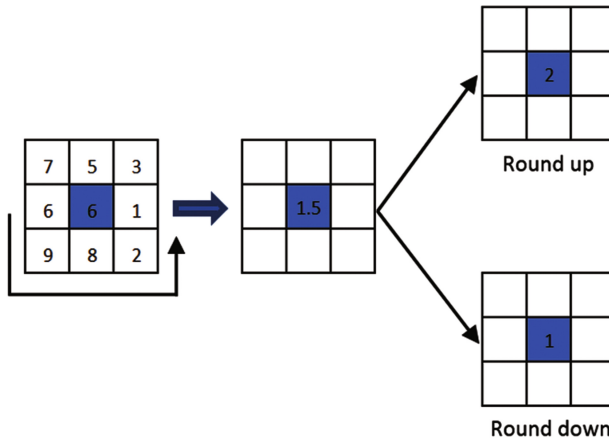


Fig. 1. Example of LIAD for a 3×3 pixel region from an image. The area of a given region is computed with $h = 0.05$. Decimal value and two nearest integers were obtained.

Table 1. The value of h parameter and the corresponding highest dimensionality of LIAD histogram

h	Maximal dimensionality	Round down	Round up
0.05	89.25	89	90
0.04	71.40	71	72
0.03	53.55	53	54
0.02	35.70	35	36
0.01	17.85	17	18

2.2 Illumination Preprocessing Method

Tan and Triggs (Tan-Triggs or TT) [18] have introduced an illumination preprocessing method for face recognition under illumination variation conditions. It uses a chain of operations such as Gamma correction, Gaussian filtering, masking, and contrast equalization, to eliminate most of changing illumination effects. For a given image I , the implementation of TT consists of three main steps as follows:

Step 1. Perform Gamma correction

$$I_g = \begin{cases} I^\gamma & \gamma > 0 \\ \log(I) & \gamma = 0 \end{cases}, \quad (5)$$

where $\gamma \in [0,1]$ is a user-defined parameter.

Step 2. Apply difference of Gaussian (DoG) [10] filtering on the obtained result at step 1.

$$D = I_g * DoG. \quad (6)$$

Step 3. Perform contrast equalization based on a three stage process as follows:

$$D = \frac{D}{(\text{mean}(|D|^\alpha))^{1/\alpha}}. \quad (7)$$

$$D = \frac{D}{(\text{mean}(\min(\tau, |D|^\alpha))^{1/\alpha}}. \quad (8)$$

$$D = \tau * \tanh\left(\frac{D}{\tau}\right). \quad (9)$$

Here, α is a strongly compressive exponent that reduces the influence of large values, τ is a threshold used to truncate large values after the first phase of normalization, and the mean is over the whole (unmasked part of the) image. The output of above steps is an image with pixel intensity in the range $(-\tau, \tau)$.

3 Face Recognition Using the TT Method and the Local Intensity Area Descriptor

In the proposed approach, face images are first normalized with the lighting components by the TT method. Next, the preprocessed images are encoded and extracted features based on the LIAD descriptor. The classification is performed using a nearest neighbor classifier with chi-square statistics [17] as dissimilarity measure. A flowchart of this approach is shown in Fig. 2.

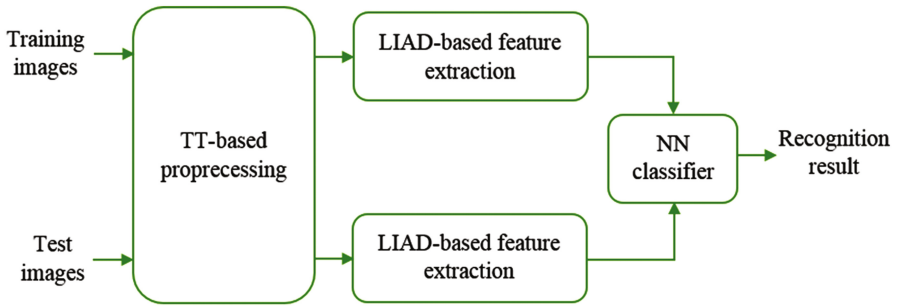


Fig. 2. Flowchart of the proposed approach for illumination variation. TT denotes the Tan-Triggs method.

4 Results and Discussions

4.1 Data Input and Experimental Settings

FERET database has 14051 grayscale images of size 256×384 pixels for 1196 individuals. It comprises of five subsets: *fa*, *fb*, *fc*, *dup I*, *dup II*. Among these, *fa* set is often used for training, *fb* set concludes expression variations, *fc* set exhibits illumination variations, and *dup I* and *dup II* sets both represent aging variations.

To test the efficiency of the proposed approach, in this study, the *fa* set was utilized for training and the *fc* set was used for testing. The face images were cropped to 150×130 pixels. To represent each image by a histogram, each image was divided into 10×10 blocks and histograms of blocks were concatenated into a single histogram. Sample images from two subsets (*fa* is in first row and *fc* is in second row) are shown in Fig. 3.

For DoG, the parameters σ_1 , σ_2 were set to 1 and 2, respectively. For GF, the parameter σ was set to 0.75. For WF, the parameters σ and α were set to 1 and 2, respectively. For SQI, the parameter σ was set to 1. For TT, the parameters σ_1 , σ_2 , and γ were set to 1, 2, and 0.2, respectively.



Fig. 3. Sample images of FERET database. First row is ten images of *fa* subset. Second row is ten images of *fc* subset.

The accuracy of the recognition method is calculated as the number of correct classifications from all classifications made. It is defined as follows:

$$\text{Accuracy (\%)} = \frac{\# \text{ of correct classifications}}{\# \text{ of total testing images}} \times 100. \quad (10)$$

4.2 Experimental Results

In this research, five well-known illumination pretreatment methods were applied and followed by five descriptors. Table 2 shows the accuracy of the proposed approach compared with some other methods using the same experimental conditions. For LDP combined with TT, the recognition accuracies with $k = 3, 4, 5$ are 69.58, 64.94, and 63.40%, respectively. Therefore, we chose the highest result that corresponded to $k = 3$ in order to compare with the other methods. For LTP combined with TT, the recognition accuracies with $t = 1, 2, 3$ are 80.92, 86.59, and 87.62%, respectively. We also chose the highest result that corresponded to $t = 1$ in order to compare with the other methods. For LIAD, a value of 0.01 is used for the h parameter, and this value obtained the highest result.

Table 2. Results of the proposed approach and the other popular methods on the FERET database

Descriptor	Lighting preprocessing method					
	NONE	DOG	GF	SQI	TT	WF
LBP	41.75	76.80	43.84	59.79	77.83	67.01
LDP	32.98	64.43	51.01	51.54	69.58	59.27
LPQ	47.93	82.44	64.43	64.94	87.11	77.83
LTP	59.27	79.38	51.54	72.68	87.62	85.56
LIAD _{up}					89.17	
LIAD _{down}					88.65	

Abbreviations: NONE: No illumination pretreatment; DOG: the difference of Gaussians method; GF: the Gradientfaces method; SQI: the self-quotient image method; TT: the Tan and Triggs method; WF: the Weber-face method.

According to the results in Table 2, the images that were preprocessed by the TT method before characteristic extraction by descriptors obtained the highest results. This implied that the TT method was the most efficient for illumination preprocessing among five methods. During the TT process, LIAD_{up}-feature extraction exhibited 1.55 to 11.34% higher than the other methods and LIAD_{down}-feature extraction was 1.03 to 10.82% higher than the other methods. These results implied that the TT followed by LIAD was more robust to illumination variation conditions.

It is known that if the dimension of feature vectors extracted from the face images is high, it will increase the time for convergence of classification and/or retrieval results. Therefore, the extracted features should be reduced in dimension to minimize time consumption for face recognition. For LBP and LPQ, they will create a histogram of size 256. For LDP with $k = 3$, a histogram of size 56 will be made. For LTP, its histogram will have a size of 512. For LIAD with $h = 0.01$, its histogram has a size of 18. This implied that the extracted feature based on LIAD has the lowest size of histogram.

In this study, a combination of TT and LIAD for face recognition under lighting variation conditions has been proposed. From the experimental results, it could be seen that not only did it achieve higher recognition rate, it also gave a lower dimensionality of the feature vector than the four existing traditional descriptors.

5 Conclusion

In this paper, an approach using the TT method to normalize the lighting of the face images and the LIAD descriptor to encode the preprocessed images has been introduced. The classification was performed by a nearest-neighbor classifier with chi-square statistic as the dissimilarity measurement. The efficiency of this approach was tested on the FERET face database. The experimental results indicated that our approach obtained higher recognition rate and a lower dimensionality of the feature vector than the LPQ, LDP, LPQ, and LTP descriptors using the same treatment to illumination variation.

Acknowledgements. This study was supported financially, in part, by grants from MOST 106-2221-E-151-010 and MOST 105-2221-E-151-010. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Tran, C.K., Tseng, C.D., Lee, T.F.: Improving the face recognition accuracy under varying illumination conditions for local binary patterns and local ternary patterns based on Weber-face and singular value decomposition. In: 2016 3rd International Conference on Green Technology and Sustainable Development (GTSD), pp. 5–9 (2016)
2. Wang, J.W., Le, N.T., Lee, J.S., Wang, C.C.: Recognition based on two separated singular value decomposition-enriched faces. *ELECTIM* **23**, 063010-1–063010-15 (2014)
3. Tran, C.K., Tseng, C.D., Shieh, C.S., Lee, T.F.: Face recognition under varying illumination conditions: improving the recognition accuracy for local ternary patterns based on illumination normalization methods and singular value decomposition. *J. Inf. Hiding Multimedia Signal Process.* **8**, 957–966 (2017)
4. Han, H., Shan, S., Chen, X., Gao, W.: A comparative study on illumination preprocessing in face recognition. *Pattern Recogn.* **46**, 1691–1699 (2013)
5. Wang, B., Li, W., Yang, W., Liao, Q.: Illumination normalization based on Weber's law with application to face recognition. *IEEE Signal Process. Lett.* **18**, 462–465 (2011)

6. Zhang, T., Tang, Y.Y., Fang, B., Shang, Z., Liu, X.: Face recognition under varying illumination using gradientfaces. *IEEE Trans. Image Process.* **18**, 2599–2606 (2009)
7. Gonzalez, R.C., Woods, R.E.: *Digital Image Processing*, 3rd edn. Prentice-Hall Inc., Upper Saddle River (2006)
8. Adini, Y., Moses, Y., Ullman, S.: Face recognition: the problem of compensating for changes in illumination direction. *IEEE Trans. Pattern Anal. Mach. Intell.* **19**, 721–732 (1997)
9. Shan, S., Gao, W., Cao, B., Zhao, D.: Illumination normalization for robust face recognition against varying lighting conditions. In: *IEEE International Workshop on Analysis and Modeling of Faces and Gestures, AMFG 2003*, pp. 157–164 (2003)
10. Marr, D., Hildreth, E.: Theory of edge detection. *Proc. Roy. Soc. Lond. Ser. B Biol. Sci.* **207**, 187–217 (1980)
11. Wang, S., Li, W., Wang, Y., Jiang, Y., Shan, J., Zhao, R.: An improved difference of Gaussian filter in face recognition. *J. Multimedia* **7**, 429–433 (2012)
12. Wang, H., Li, S.Z., Wang, Y., Zhang, J.: Self quotient image for face recognition In: *Proceedings of the International Conference on Pattern Recognition*, pp. 1397–1400 (2004)
13. Xie, X., Lam, K.M.: An efficient illumination normalization method for face recognition. *Pattern Recogn. Lett.* **27**, 609–617 (2006)
14. Ahonen, T., Hadid, A., Pietikäinen, M.: Face recognition with local binary patterns. In: Pajdla, T., Matas, J. (eds.) *Computer Vision - ECCV 2004*. LNCS, vol. 3021, pp. 469–481. Springer, Heidelberg (2004)
15. Ojala, T., Pietikäinen, M., Harwood, D.: A comparative study of texture measures with classification based on feature distributions. *Pattern Recogn.* **29**, 51–59 (1996)
16. Matti, P., Abdenour, H., Guoying, Z., Timo, A.: *Computer Vision Using Local Binary Patterns*. Computational Imaging and Vision, vol. 40. Springer, Heidelberg (2011)
17. Tran, C.K., Tseng, C.D., Chao, P.J., Ting, H.M., Chang, L., Huang, Y.J., Lee, T.F.: Local intensity area descriptor for facial recognition in ideal and noise conditions. *J. Electron. Imaging* **26**, 023011-1–023011-10 (2017)
18. Tan, X., Triggs, B.: Enhanced local texture feature sets for face recognition under difficult lighting conditions. *IEEE Trans. Image Process.* **19**, 1635–1650 (2010)
19. Ojansivu, V., Heikkilä, J.: Blur insensitive texture classification using local phase quantization. In: Elmoataz, A., Lezoray, O., Nouboud, F., Mammass, D. (eds.) *Image and Signal Processing*. LNCS, vol. 5099, pp. 236–243. Springer, Heidelberg (2008)
20. Zhou, S.R., Yin, J.P., Zhang, J.M.: Local binary pattern (LBP) and local phase quantization (LPQ) based on Gabor filter for face representation. *Neurocomputing* **116**, 260–264 (2013)
21. Jabid, T., Kabir, M.H., Chae, O.: Local directional pattern (LDP) for face recognition. *Int. J. Innov. Comput. Inf. Control* **8**, 2423–2437 (2012)
22. Dalal, N., Triggs B.: Histograms of oriented gradients for human detection. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005*, vol. 881, pp. 886–893 (2005)
23. Phillips, P.J., Moon, H., Rizvi, S.A., Rauss, P.J.: The FERET evaluation methodology for face-recognition algorithms. *IEEE Trans. Pattern Anal. Mach. Intell.* **22**, 1090–1104 (2000)
24. Kreyszig, E.: *Advanced Engineering Mathematics*, 10th edn. Wiley, New York (2010)

Analysis of the Dynamic Co-purchase Network Based on Image Shape Feature

Xiaoyin Li^(✉) and Jean Lai^(✉)

Department of Computer Science, Hong Kong Baptist University,
Kowloon, Hong Kong
16439929@life.hkbu.edu.hk, Jeanlai@comp.hkbu.edu.hk

Abstract. E-commerce has become popular and profitable because it provides lots of convenience for both retailers and buyers. We introduce an idea that co-purchase network may have correlation with image shape feature. We develop a simple image shape retrieval model to automatically identify the items that the buyer would be interested. The results provide new insights into user behavior. Based on the results, we found that people would buy items which are similar in shape with the item that they have already selected. Based on this finding, it can provide more personalized recommendation of co-purchase items to the customer.

Keywords: Co-purchase · Image feature · User behavior

1 Introduction

E-commerce has become popular and profitable because it provides lots of convenience for both retailers and buyers. Retailers can easily track buyer's behavior according to the sale figures captured from the e-commerce shop. Vital marketing is using social networking services to spread their self-replicating viral processes to achieve marketing goals or benefit. E-Commerce websites proactively recommend products to their target customers. In order to improve the recommendation effectiveness, the authors of [4] built a system by analysing clothing attributes (such as color) to provide occasion-oriented recommendation. In [5], the authors propose and approach to segment the clothing regions to detect the type of clothing. Then based on detected result, retrieve visual similarity image from database. In [6], the authors propose a recommend system based on images' styles and relationship. In this report, we focus on content-based image information by using image feature extraction, such as shape, based on Amazon product provided by [7]. There is metadata sample present for a record in Table 1. Comparing the co-purchase items image to find out the similarity between them through their shape.

Table 1. Sample Amazon product metadata.

Property	Value
asin	0000031852
title	Girls Ballet Tutu
price	3.17
imUrl	http://ecx.images- amazon.com/images/I/51fAmVkTbyL._SY300_.jpg
related	also_bought ["B00JHONN1S"]
	also_viewed ["B002BZX8Z6"]
	bought_together ["B002BZX8Z6"]
saleRank	{"Toys & Games": 211836 }
brand	Coxlures
categories	[["Sports & Outdoors", "Other Sports", "Dance"]]

2 Prior Work

There has a great deal of work related to the co-purchase recommendation or content-based image retrieval. However, there is no one combined them together to see whether they would affect each other or not. According to the subject, I not only need to find the methods to extract the features of image, but also to find some suitable methods to estimate the similarity of shape and color.

2.1 Cosine Similarity

Cosine similarity is one of the methods to measure two non-zero vectors distance, as illustrated in Fig. 1. The cosine of the angle between this two vectors represent how similarity they are. The cosine value is within -1 to 1 . Cosine value closer to 1 , the more similarity they are. Cosine value closer to -1 , the more dissimilarity they are. The Eq. (1) [11] shows how to calculate the cosine similarity.

$$similarity = \cos \theta = \frac{A \cdot B}{\|A\|_2 \|B\|_2} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}} \quad (1)$$

where A and B are two vectors. A_i and B_i are the components of A and B .

2.2 Pearson Correlation Coefficient

The Pearson correlation coefficient is used to measure linear relationships between two datasets. The greater the absolute value of the correlation coefficient, the stronger linear correlation between these two datasets. The absolute value of correlation coefficient closer to 0 , the weaker linear correlation between these two datasets. The Eq. (2) [12] shows how to calculate the Pearson correlation coefficient.

$$r = r_{xy} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \sqrt{n \sum y_i^2 - (\sum y_i)^2}} \quad (2)$$

where n is the number of components in a dataset, x_i and y_i are the components of X and Y datasets respectively, r is the coefficient between X and Y datasets.

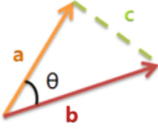


Fig. 1. Cosine similarity.

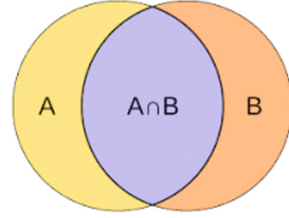


Fig. 2. Jaccard similarity coefficient.

2.3 Jaccard Similarity Coefficient

Jaccard similarity coefficient is another way to compare the similarities and differences between two datasets. The value of Jaccard similarity coefficient closer to 1, the higher similarity between these two datasets. The two datasets A and B given as illustrated in Fig. 2. According to the Jaccard similarity coefficient Eq. (3) [10],

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| + |A \cap B|} \quad (3)$$

where A and B are two datasets. $A \cap B$ is the intersection of two datasets. $A \cup B$ is union of two datasets. If A and B datasets are both empty, we define $J(A, B) = 1$.

3 Image Processing

There are thousands of images in different color and shape. To have better analysis result, we enhance the images as a pre-processing step before calculating the similarity. Due to the most of Amazon products background are in white, our methods are based on the assumption that the item’s image are white background. Figure 3 is an example of the original image and corresponding co-purchase item image that retrieve from dataset in [7].

3.1 Pre-processing

There are few steps of shape pre-processing: (1) **Convert the image into 8-bit grayscale image.** Because we only need the shape information of the image, we



Fig. 3. The original image (left) and corresponding co-purchase item image (right).

discard the color information by converting the image into 8-bit grayscale. (2) **Smooth the image.**

This step is smooth the edges of the shape by using average mask to obtain high connectivity of the boundary. (3) **Enhance contrast between background and foreground.** We enhance the contrast between background and its foreground by equalizing the histogram. Foreground can be also called the shape of image. As long as there is contrast between background and its shape, the shape can be said to exist, and can then be detected [8]. (4) **Binary threshold segmentation.** Get the binary image by using threshold value to segment the shape and its background. In this case, we set the threshold value equals 220. It means that the pixel value larger or equal to 220, it would assign to background part. The pixel value smaller than 220, it would assign to the shape part. Each step processed result image as illustrated in Fig. 4.

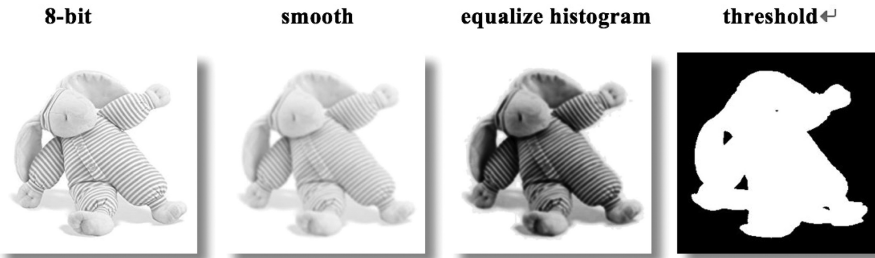


Fig. 4. Shape pre-processing steps.

3.2 Shape Similarity

Before calculating the shape similarity, there are few steps need to be settle down. First, cropping the images after selecting the region of interest (ROI) according to the position of the leftmost, rightmost, topmost and lowermost point. Then, linear re-scaling two images so they have the same size. The result as illustrated in Fig. 5.

Then, we divide the image in half: the left side and the right side. Store the x-coordinate value of the boundary from top to the bottom, from left side to right side. The line chart as demonstrated the correlation relationship between two datasets in Figs. 6 and 7. The X-axis is represented the count value of dataset. The y-axis is



Fig. 5. Re-scaling result of image processing.

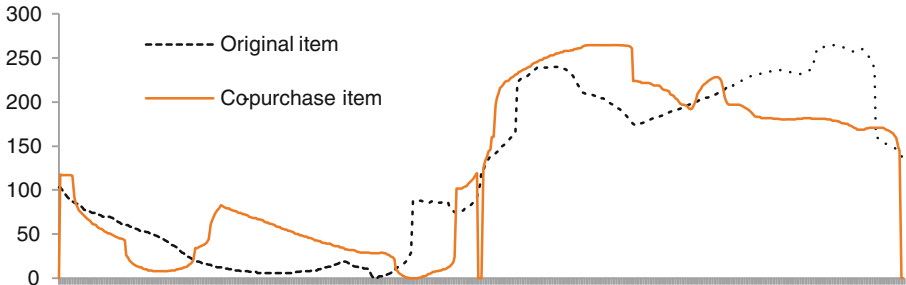


Fig. 6. The x-coordinate value of the boundary in two images (Fig. 3). The X-axis is represented the count value of dataset. The y-axis is represented the x-coordinate value at the boundary in the image.

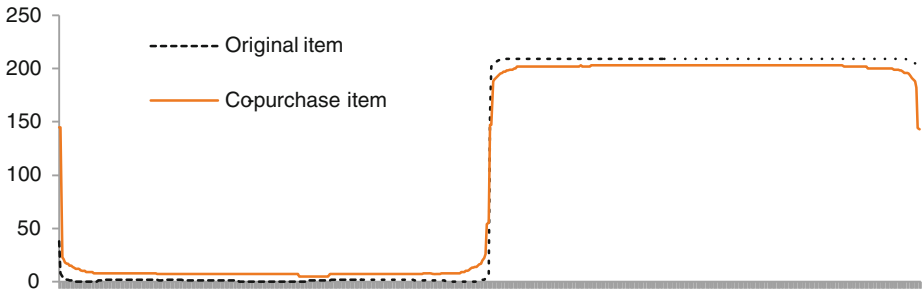


Fig. 7. The x-coordinate value of the boundary in two images (Fig. 8). The X-axis is represented the count value of dataset. The y-axis is represented the x-coordinate value at the boundary in the image.

represented the x-coordinate value at the boundary in the image. After store two datasets, calculate the similarity by using Pearson correlation coefficient and cosine similarity.

However, the angle of the shape may have great influence the result of similarity. The example as show in Fig. 9. So, the angle of the shape should be considered.



Fig. 8. The shape of co-purchase items images which closer to the rectangle.



Fig. 9. Different angle of shape.

Our solution is keep an image fixed, rotate another image 1 degree per times until rotate 360°. For each time rotation, cropping the images after selecting the region of interest (ROI) according to the position of the new leftmost, rightmost, topmost and lowermost point. Then, linear re-scaling two images so they have the same size. Then calculate the Pearson correlation coefficient and cosine similarity until find the maximum similarity value respectively.

4 Experimental Result

In this section, we mainly display the result of shape similarity between different images.

4.1 Shape Similarity Result

Table 2 is the shape similarity result. According to the Table 2, the most pair of images gain high Pearson correlation coefficient result, which means they have strong similarity. But the truth is that some of them may not be seen as high similarity at shape based on our human visual system. Same as the Cosine similarity, we think if both the shape of images closer to the rectangle, the higher similarity will obtain.

On the contrary, Jaccard result seems more fit our human visual system. The index from (a) to (g) are the co-purchase items retrieve from [7]. The index from (h) to (l) image set are choose by randomly from image dataset, which as illustrated in Table 3.

Table 2. Table captions should be placed above the tables.










Image Set	Index	Pearson	Cosine	Jaccard
	(a)	0.94	0.97	0.83
	(b)	0.98	0.99	0.79
	(c)	0.99	0.99	0.66
	(d)	0.94	0.97	0.65
	(e)	0.85	0.94	0.62
	(f)	0.84	0.92	0.61
	(g)	0.91	0.96	0.51

Table 3. Co-purchase and non-co-purchase items' shape similarity results.

Image Set	Index	Pearson	Cosine	Jaccard
	(b)	0.98	0.99	0.79
	(h)	0.90	0.97	0.58
	(i)	0.79	0.94	0.45
	(j)	0.77	0.92	0.39
	(k)	0.77	0.92	0.38
	(l)	0.92	0.91	0.30

5 Conclusion

There has a great deal of work related to the co-purchase recommendation or content-based image retrieval. However, there is no one combined them together to see whether they would affect each other or not. According to the Table 2, for those

co-purchase items, their Jaccard similarity usually get high result. According to the Table 3, based on their Jaccard similarity result, co-purchase items are higher than non-co-purchase items. Therefore, we believe that human select co-purchase items may affect by image shape feature. Moreover, based on this finding, it provides more personalized recommendation of co-purchase items to the customer.

6 Further Work

The methods we used in this study are the basic methods and our data was limited at Amazon. The accuracy of feature extract might be low. In this study, we are not considered the situation that two objects are in one image. Also, we rotate the image left and right but not forward and backward. For the further work, we will try to settle down these limitations. Using the questionnaire survey method to verify and improve the result. At the meantime, find the more suitable method to do the feature extract and recognition.

References

1. Leskovec, J., Adamic, L., Huberman, B.A.: The dynamics of viral marketing. *ACM Trans. Web* **1**(1), Article 5, May 2007. 39 pages. doi:[10.1145/1232722.1232727](https://doi.org/10.1145/1232722.1232727)
2. Basuchowdhuri, P., Shekhawat, M.K., Saha, S.K.: Analysis of product purchase patterns in a co-purchase network (2014). doi:[10.1109/EAIT.2014.11](https://doi.org/10.1109/EAIT.2014.11)
3. Linden, G., Smith, B., York, J.: Amazon.com recommendation - item-to-item collaborative filtering. *IEEE Internet Comput.* **7**(1), 76–80 (2003)
4. Liu, S., Feng, J., Song, Z., Zhang, T., Lu, H., Xu, C., et al.: Hi, magic closet, tell me what to wear. *ACM* (2012)
5. Kalantidis, Y., Kennedy, L., Li, L.-J.: Getting the look: clothing recognition and segmentation for automatic product suggestions in everyday photos. *ACM* (2013). doi:[10.1145/2461466.2461485](https://doi.org/10.1145/2461466.2461485)
6. McAuley, J., Targett, C., Shi, Q., Hengel, A.V.D.: Image-based recommendations on styles and substitutes (2015)
7. Amazon product data. <http://jmcauley.ucsd.edu/data/amazon/links.html>. Accessed 31 May 2017
8. Nixon, M.S., Aguado, A.S.: Feature Extraction and Image Processing. Academic, Amsterdam, London, Boston (2008)
9. Euclidean distance, Wikipedia. https://en.wikipedia.org/wiki/Euclidean_distance. Accessed 31 May 2017
10. Jaccard index, Wikipedia. https://en.wikipedia.org/wiki/Jaccard_index. Accessed 31 May 2017
11. Cosine Similarity, Wikipedia. https://en.wikipedia.org/wiki/Cosine_similarity. Accessed 31 May 2017
12. Pearson correlation coefficient, Wikipedia. https://en.wikipedia.org/wiki/Pearson_correlation_coefficient. Accessed 31 May 2017

VQ Compression Enhancer with Huffman Coding

Chin-Feng Lee^{1(✉)}, Chin-Chen Chang², and Qun-Feng Zeng³

¹ Chaoyang University of Technology, No. 168, Jifeng E. Rd.,
Wufeng District, Taichung 41349, Taiwan, R.O.C.

lcf@cyut.edu.tw

² Feng Chia University, No. 100, Wenhwa Rd., Seatwen,
Taichung 40724, Taiwan, R.O.C.

alan3c@gmail.com

³ National Chung Cheng University, No.168, Sec. 1, University Rd.,
Minhsiung, Chiayi 62102, Taiwan, R.O.C.

qf801002@gmail.com

Abstract. Vector quantization (VQ) is an effective and important compression technique with high compression efficiency and widely used in many multimedia applications. VQ compression is a fixed-length algorithm for image block coding. In this paper, we employ the Huffman Coding technology to enhance VQ compression rate and get a better compression performance due to the reversibility of the Huffman Coding. The proposed method exploits the correlation between neighboring VQ indices with similarity. The similarity draws a large number of small differences from the current index with that of its adjacent neighbors; thereby, increasing the compression ratio due to the great quantity of small differences. The experimental results reveal that the proposed combination technique adaptively provides better compression ratios at high compression gains than that of VQ compression. The proposed method is superior in smoother pictures with the compression gains greater than 100%; even for the complex images the compression gain can be increased more than 25%. Therefore, the VQ-Huffman method can really enhance the efficiency of VQ compression.

Keywords: Vector quantization (VQ) · Huffman coding · Compression ratio

1 Introduction

The development of scientific and technological information bringing computer into family, computer becomes a part of human life audio-visual entertainment. But the current development of software design requires a huge amount of hard disk space. Moreover, with the progress of digital technology and the vigorous development of the Internet, a variety of including image, video, audio and other multimedia application innovation. As multimedia data storage needs increased, the data compression is more important and cannot be avoided. Good compression ability in data compression not only decreases the storage space but also great reduces transmission time. About data transmission security, since the data have been compressed and processed, the

information is not the original format, can only restore by decompressed operations while greatly increasing the information security.

Multimedia compression is divided into two categories: one is lossless compression also called reversal compression, which allows the original data to be perfectly reconstructed from the compressed data. This type of compression is generally used for text files, executive files, medical image and important information. The other is lossy compression also called irreversible compression. The reconstructed results of such data are not exactly the same as the data before they are being processed. But most of the information and features of the original data are still included in the results of the recovered data. Lossy compression has a much more compression rate than lossless compression; therefore, lossy compression is often used for image compression (such as VQ compression, JPEG compression and MPEG compression) and sound compression (such as MP3) which require high compression rate and allow partial distortion.

Vector quantization (VQ) is an effective and important compression technique, which can achieve very high compression efficiency; therefore VQ is widely used in many different situations, such as multimedia systems, high-definition television, telex systems, image data base management, and so on. VQ compression is a lossy data compression method based on the principle of block coding and is a fixed-to-fixed length algorithm. We use VQ compression as a basis due to the high compression rate and fast implementation of VQ. Then we employ the Huffman Coding technology to enhance VQ compression rate and get a better compression performance due to the reversibility of the Huffman Coding. The proposed VQ-Huffman method can exactly restore the VQ index table.

VQ compression is often used in multimedia including, sounds, images, text [1] and so on. Also VQ compression can add to the information hiding, for example, combined with Search ordering coding (SOC) [2, 3] or locally adaptive coding (LAC) [4, 5]. If we can have a better compression ratio on the basis of VQ, then we can transmit faster, also there will be more space to hide more secret data. This paper uses the difference between current index and surrounding index [6, 7], and then use Huffman Coding for further compression, hoping to achieve better compression and enhance VQ compression efficacy.

The rest of this paper is organized as follows. Section 2 introduces the related work. The proposed VQ-Huffman method is presented in Sect. 3. The experimental results are shown in Sect. 4. Conclusions make recommendations for applying the results in the last section.

2 The Related Work

Image compression using vector quantization is introduced in Sect. 2.1 and then the principle method of Huffman coding is summarized in Sect. 2.2.

2.1 Vector Quantization

In 1980, Linde, Buzo, and Gray (LBG) proposed a VQ algorithm to design a good codebook for data compression. A vector quantizer is composed of two parts, encoder

and decoder. An encoder will compare each input vector with every codevector (also called a codeword) in the codebook and generate the index of best-matched codeword which has the minimum distortion between the input vector and the codevector.

The decoded of VQ is to reconstruct the vector after quantization. A decoder takes the index corresponding to each image block to search the codebook and to locate the codevector in that codebook and generates the output vector. We repeat the decoding steps until every index is processed, and the collection of all the result vectors can be used to recovery an image similar to the original one by the quantization reconstruction.

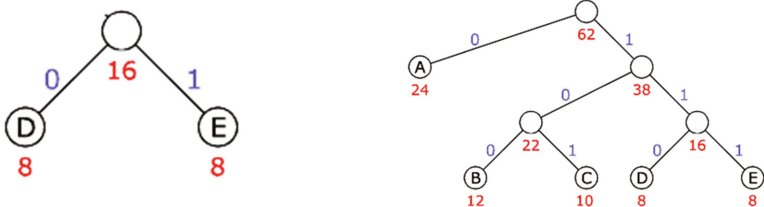
2.2 Huffman Coding

In 1952, Huffman proposed an encoding algorithm [9] which permits error-free data reconstruction. Huffman coding represents data by variable-length codes, with more frequent data being represented by shorter codes.

We use a simple example to show how a Huffman's Tree is built. Let the character set $Sym = (A, B, C, D, E)$ for corresponding frequency set of symbol occurrences $Freq = (24, 12, 10, 8, 8)$. The technique works by creating a binary tree of nodes. Initially, all nodes are leaf nodes, which contain the symbol itself and the frequency of appearance of the symbol.

The process essentially begins with a new node whose children are the 2 nodes with smallest frequencies, such that the frequency of new internal node is equal to the sum of the children's frequency as shown in Fig. 1(a). As a common convention, bit "0" represents following the left child and bit "1" represents following the right child. With the previous 2 nodes merged into one internal node and with the new node being now considered, the procedure is repeated until only one node remains as shown in Fig. 1(b), the Huffman tree. Finally, the symbol A will be encoded by "0"; the symbol B will be encoded by "100"; the symbol C will be encoded by "101"; the symbol D will be encoded by "110", and the symbol E will be encoded by "111", respectively.

Huffman constructs a binary code tree from the bottom up, that omits unused symbols to produce the most optimal code lengths. Though the codes are of different bit lengths, they can be uniquely decoded.



(a) An internal node with two leaf nodes

(b) A Huffman's tree

Fig. 1. The process of a Huffman's tree

3 The Proposed Method

Each natural image mostly possesses the local spatial correlation which makes the neighboring positions in image pixels have the similar intensity. When we train a codebook by using VQ technique, we also sort the codewords $CW_i = (cw_{i1}, cw_{i2}, \dots, cw_{ik})$ according to S_i in increasing order, where $S_i = \sum_{j=1}^k cw_{ij}$. Afterwards, we encode an input image by the sorted codebook and we can exploit the correlation between neighboring VQ indices with similarity. The similarity draws a large number of small differences from the current index with that of its adjacent neighbors; thereby, increasing the compression rate due to the great quantity of small differences.

The seed blocks are at the first column and the first row and the other blocks are called residual blocks as shown in Fig. 2. Indices of the seed blocks are kept unshorten but the indices for residual blocks are encoded with shorter bit streams. Figure 3 displays the layout of the current to-be-processed block X as well as its upper and left neighbors, U and L respectively. The encoding procedure will encode each residual block by two parts. The first part called leading indicators (LI) and the second part is codes for the difference values between the indices of (X and U) or (X and L). Exploiting Huffman encoding, there will be seven cases generated as shown in Fig. 4. The VQ-Huffman encoding with reversibility can further help how to encode each residual block and enhance the compression efficiency of VQ compression.

Case 1: The leading indicator is set as LI = “00” if the indices of upper block U and current block X are the same.

While the indices of upper block U and current block X are different, i.e., the difference value $Ue \neq 0$, we will go the following six cases.

Case 2: The leading indicator is set as LI = “01” when the different value $Le = 0$ because the left block L and the current block X have same index.

Case 3: While $Ue \neq 0$ and $1 \leq |Le| \leq 4$, the leading indicator is set as “100”. Since there are eight difference values, we encode each difference with 3 bits such that the difference 1 can be encoded as “001”, 2 as “010”, 3 as “011”, 4 as “100”, -1 as “101”, -2 as “110”, -3 as “111”, -4 as “000”.

Case 4: While $Ue \neq 0$ and $5 \leq |Le| \leq 8$, the leading indicator is set as “101”. There also exists eight differences so we encode the difference 5 as “001”, 6 as “010”, 7 as “011”, 8 as “100”, -5 as “101”, -6 as “110”, -7 as “111”, -8 as “000”.

Case 5: While $Ue \neq 0$ and $9 \leq |Le| \leq 12$, we set the leading indicator LI = “110”. With nine difference values in this case, we will encode the difference 9 as “001”, 10 as “010”, and so on.

Case 6 has the condition $Ue \neq 0$ and $13 \leq |Le| \leq 28$ while Case 7 has the condition that $Ue \neq 0$ and $|Le| \geq 29$. Case 6 sets the leading indicator LI = “1110” while Case 7 has LI = “1111”.

In Case 6, there are 32 difference values, that is to say, the difference values from 13 to 28 and -13 to -28 are encoded by “00001”, “00010”, “00011”, and so on. As for

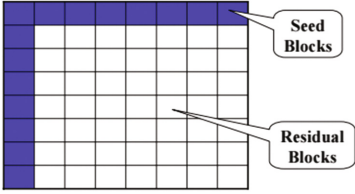


Fig. 2. Seed blocks and residual blocks

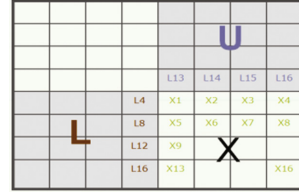


Fig. 3. Current Block X , upper block U and left block L

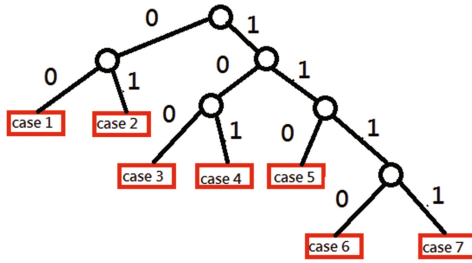


Fig. 4. Seven cases of VQ-Huffman Coding

the seventh case, we directly encode the current block X with the leading indicator and its own index value in binary representation.

4 Experimental Results

We used eight standard grayscale test images of size $H \times W$ where H and W respectively stand for the height and width of an image; $H = W = 512$ as shown in Fig. 5 for our experiment. Compression Ratio (CR) [10] is a very logical way of measuring how well a compression algorithm compresses a given digital image. The ratio of the number of bits required to represent the image before compression into the number of bits required to represent the image after compression.

In the experiment, the CR is computed by dividing the size of the original image block in bits over the size of the compressed image block. Thus Eq. 1 gives the VQ compression ratio because each grayscale image block has $h \times w$ pixels and each pixel uses 8 bits to represent it. Therefore, L_x is constant because each image block is encoded by an 8-bit index using VQ compression. So the compression ratio by VQ is $(8 \times 16)/8 = 16$.

$$CR = (8 \times h \times w)/L_x \tag{1}$$

From Eq. 1, we know that the smaller L_x the larger CR. That is the reason why we exploit Huffman Coding to lower the size of the compressed image block, for creating



Fig. 5. Eight standard grayscale test images, size 512×512 pixels

higher compression ratio. The compression ratio of proposed VQ-Huffman method can be redefined as Eq. 2 below. According to Eq. 2 the CR will become higher and so does the compression effect.

$$CR = \frac{8 \times h \times w}{L_v}, \text{ where } L_v = \sum_{i=1}^{\frac{h}{8} \times \frac{w}{8}} L_i \quad (2)$$

Table 1 provides the performance measured by VQ compression in comparison with the proposed VQ-Huffman method. Our proposed method has the average compression ratio of up to 24.857 for eight grayscale images. For the smoother pictures such as Tiffany and Toys, the compression gains can reach greater than 100%; even for the complex images such as Lena, Peppers, Gold, Barbara and F-16, the compression gain can be increased more than 25%. The results of Table 1 show that our method has a significant lift in compression gains.

Table 1. Comparisons of different encoding techniques with codebook size 256

Compression rate	VQ method	Propose method	Compression gain
	(A)	(B)	(B - A)/A (%)
Toys	16.0	33.336	108.35
Tiffany	16.0	32.520	103.25
F-16	16.0	25.472	59.20
Peppers	16.0	23.320	45.75
Lena	16.0	22.096	38.10
Girl	16.0	21.112	31.95
Gold	16.0	21.000	31.25
Barbara	16.0	20.000	25.00
Avg.	16.0	24.857	55.36

Table 2 displays the number of blocks after VQ-Huffman method is conducted to each test image. As mentioned earlier, each image block can be encoded into a 8-bit index by VQ compression. Rather than fixed-length encoding, the proposed method encoded each image block into 2 bits by Cases 1 and 2; 6 bits by Cases 3, 4, and 5; 9 bits by Cases 6 and 12 bits by Cases 7. Therefore, we exploit our method to encode the image block by first, two cases, or else try to make the block encoded in the third, fourth, fifth cases so as to really enhance the compression ratios. From Table 2, most blocks are indeed encoded by Cases 1 or 2, which makes the VQ-Huffman method superior to that of VQ compression.

Table 2. Number of blocks for each case by VQ-Huffman method

# of blocks	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7
Toys	8267	2931	1934	618	451	856	1072
Tiffany	7793	1900	3431	1284	589	792	340
F-16	5292	3061	2743	996	582	1287	2168
Peppers	4945	2145	1993	1799	1105	1835	2307
Lena	5019	1444	2107	1768	1143	1891	2757
Girl	3863	1644	2178	1977	1537	2380	2550
Gold	3026	2328	2217	1950	1461	2765	2382
Barbara	3540	1905	1861	1724	1182	2568	3349

5 Conclusions

Our paper proposed a combination of VQ compression with Huffman coding compression for enhancing image compression and exactly restoring to the VQ index table. Experimental results show that the proposed combination technique adaptively provides better compression ratios at high compression gains than that of VQ compression.

The overall superior performance offered by our method encourages the method to be applied directly in some applications such as digital libraries or online multimedia transmission. Moreover, it also can be applied to some interdisciplinary research topics such as information hiding to help prevent unauthorized access, use, disclosure, disruption, modification, or inspection of information. Since the proposed method contributes to completely restore the VQ index table without data loss, the VQ-Huffman method can provide to embed a large amount of data into VQ images for information hiding applications.

Acknowledgments. This research was partially supported by the Ministry of Science and Technology of the Republic of China under the Grants MOST 105-2221-E-324-014 and MOST 105-2221-E-035-051.

References

1. Korycki, R.: Authenticity examination of compressed audio recordings using detection of multiple compression and encoders' identification. *Forensic Sci. Int.* **238**, 33–46 (2014)
2. Chang, C.C., Chen, G.M., Lin, M.H.: Information hiding based on search-order coding for VQ indices. *Pattern Recogn. Lett.* **25**, 1253–1261 (2004)
3. Lin, C.C., Liu, X.L., Yuan, S.M.: Reversible data hiding for VQ-compressed images based on search-order coding and state-codebook mapping. *Inf. Sci.* **293**, 314–326 (2015)
4. Yang, C.H., Lin, Y.C.: Fractal curves to improve the reversible data embedding for VQ-indexes based on locally adaptive coding. *J. Vis. Commun. Image Represent.* **21**, 334–342 (2010)
5. Chang, C.C., Nguyen, T.S., Lin, C.C.: A reversible data hiding scheme for VQ indices using locally adaptive coding. *J. Vis. Commun. Image Represent.* **22**, 664–672 (2011)
6. Chang, C.C., Kieu, T.D., Wu, W.C.: A lossless data embedding technique by joint neighboring coding. *Pattern Recogn.* **42**, 1597–1603 (2009)
7. Wang, J.X., Lu, Z.M.: A path optional lossless data hiding scheme based on VQ joint neighboring coding. *Inf. Sci.* **179**, 3332–3348 (2009)
8. Linde, Y., Buzo, A., Gray, R.M.: An algorithm for vector quantizer design. *IEEE Trans. Commun.* 702–710 (1980)
9. Huffman, D.A.: A method for the construction of minimum-redundancy codes. *Proc. Inst. Radio Eng.* **40**(9), 1098–1101 (1952)
10. Telagarapu, P., Naveen, V.J., Prasanthi, A.L., Santhi, G.V.: Image compression using DCT and wavelet transformations. *Int. J. Signal Process. Image Process. Pattern Recogn.* **4**, 61–70 (2011)

Adaptive Steganography Method Based on Two Tiers Pixel Value Differencing

Chi-Yao Weng¹, Yen-Chia Huang¹, Chin-Feng Lee²(✉),
and Dong-Peng Lin²

¹ Department of Computer Science, National Pingtung University,
90003 Pingtung, Taiwan, R.O.C.

cyweng@mail.nptu.edu.tw, q7273617@yahoo.com.tw

² Department of Information Management, Chaoyang University of Technology,
41349 Taichung, Taiwan, R.O.C.

lcf@cyut.edu.tw, monkey60121@gmail.com

Abstract. The pixel value differencing (PVD) scheme provided high embedding payload with imperceptibility in the stego images. In their approach, they used two pixels differencing to represent the complexity of pixels, and applied it to estimate how many bit will be hidden into. As the difference with small value, it means that two pixels can not tolerated with larger change, therefore, few secret bit should be embedded into these pixels. PVD scheme did not completely take pixel tolerance into consideration because of only applying one criterion, pixel differencing. In this paper, a new data hiding scheme using PVD operation and incorporating with pixel tolerance into a cover image is proposed. The pixel tolerance indicates that a greater pixel-value is more change of gray-value could be tolerated. Following up this idea, our proposed scheme applies a threshold (TH) and two quantization tables to hide secret data into a block with two pixels using modified k -bits LSB. The number of k -bits is adaptive and depends on the quantization tables setting. The adjustment strategy is to maintain the differencing value in the same range. The experimental results show that our scheme is superior to those in the previous literature.

Keywords: Pixel value differencing (PVD) · Pixel tolerance · LSB substitution · Adaptive data hiding · Adjustment strategy

1 Introduction

Recently, data hiding techniques (or called as steganography) have become an important research issues because it can enhance the security of network communication. For the purpose of network communication, two schemes are adopted to protect secret message during data transmission. One is data encryption, where the data is encoded by using a secret key before data sending. As receiving encryption data, receiver uses the same secret key to decode the data. The most popular encryption techniques [1] are DES, AES, RSA, and so on. The other way is data hiding, where the message is hidden into a cover media. The cover media might be a text, a video, a map, and so on [2, 3].

Numerous data hiding techniques have been developed [4–16], and they often recommended a digital image as a cover media. The most simplest and common data hiding method is Least-significant-bit (LSB). In this manner, the secret k bits are replaced each pixel with number of k LSB position, k -bits LSB for shorting [5]. LSB method suffered huge image distortion since k is larger. The range of image distortion using LSB method is $[0, 2^{k-1}]$. For decreasing image distortion, the Modified LSB is thus proposed. In the Modified LSB, it used one bit of MSB, which the position of MSB is $k + 1$ position, to adjust the distortion of k -bits LSB and let the distortion is smaller. For example, $k = 3$, Modified LSB used bit of MSB-4 ($= 3 + 1$) to alter the pixel value. The range of image distortion in Modified LSB approach is $[0, 4](= [0, 2^{k-1}])$. It is obviously known that Modified LSB approach has better image quality than LSB, $[0, 2^{k-1}] < [0, 2^{k-1}]$. Although each pixel in a cover image had the same distortion, $[0, 2^{k-1}]$ or $[0, 2^{k-1}]$, they did not take pixel complexity into consideration for human vision system. The pixel complexity means that the pixel falling into edge area could be tolerated with large changed than that of pixel falling into smooth area. The pixel value differencing method is thus proposed for achieving pixel complexity.

Pixel value differencing (PVD) approach is generated by Wu and Tsai [6]. They used the differencing between two consecutive pixels to determine the pixel complexity, and to decide how many bits would be hidden into cover blocks with two consecutive pixels. PVD method is good technique for solving the drawback of LSB, but the hidden capacity is fewer. Yang and Weng proposed a multi-pixel differencing (MPD) method to enrich hidden capacity [7]. In their approach, they found out multi-group within a block with four pixels and let groups to realize PVD concept. Wu et al. exploited the concept of PVD and proposed a new data hiding based on PVD and LSB [8]. Their method used the pixel differencing to judge which one manner, PVD or LSB, would be applied into two consecutive pixels. Yang et al. proposed an adaptive data hiding method in edge area based on LSB substitution [12]. Their approach conceals more secret bits in edge area than that of in the smooth area. Wang et al. created high image quality strategy based on PVD and modulo function [13]. So far, several proposed schemes still utilize PVD concept to promote the performance in terms of high image quality [9, 13] and larger embedding capacity [7, 8, 10–18].

The rest of this paper is organized as follows. Section 2 briefly reviews Wu and Tsai's pixel value differencing method [6], and Wu et al.'s pixel value differencing and LSB approach [8]. The proposed method of adaptive data hiding, including the embedding and extracting procedures, describes in Section 3. Section 4 presents the experimental results and compares the performance of the presented approach with that of a previously proposed scheme. Section 5 draws some conclusions.

2 Literature Review

This section will briefly review two proposed schemes, which is related to the work of pixel value differencing and PVD and LSB substitution. The first scheme is PVD scheme developed by Wu and Tsai [6]. The second scheme is PVD and LSB substitution developed by Wu et al. [8].

2.1 Wu and Tsai's Pixel Value Differencing

The gray level image is used as a cover image in PVD method. A cover image is divided into non-overlapping blocks of two consecutive pixels, saying p_x and p_{x+1} . From each block, the pixel difference is obtained and modified as a new pixel difference for data embedding. A larger differencing value indicates that a block allows a grater modification. The data hiding algorithm is implemented as follows:

Step 1: Compute the differencing value d_i for each block of two pixels, p_x and p_{x+1} , where $d_i = |p_x - p_{x+1}|$.

Step 2: Find the optimal range R_i from a contiguous range, saying R_i where $i = 1, 2, 3, \dots, r$. The width of R_i is $u_i - l_i + 1$, where u_i is the upper bound of R_i and l_i is the lower bound of R_i .

Step 3: Decide n bits of secret data which are concealed into each differencing value d_i . Then, fetch n bits from a secret binary string and transform n bits into decimal value b . For example, assume a secret binary string is 1011, then $b = 11$.

Step 4: Obtain new pixel differencing value d'_i according to following equation.

$$d'_i = \begin{cases} l_i + b & , \text{ if } d_i \geq 0 \\ -(l_i + b) & , \text{ if } d_i < 0 \end{cases} \quad (1)$$

Step 5: An inverse calculation from d'_i is performed to yield the new gray value of two pixels in a block, saying p'_x and p'_{x+1} .

Step 6: Repeat Steps 1–5, until all secret data are hidden into cover image or all non-overlapping blocks are processed, and the stego image is obtained.

In the extraction phase, the secret data are extracted from each blocks of stego image is the same order as the embedding phase. The information of contiguous range is necessary to find optimal range R_i and to determine how many secret bits are to be hidden into two pixels. In addition, the embedded bits in a block can easily to be extracted by subtracting the lower bound of R_i from the differencing between two consecutive pixels.

2.2 Wu et al. PVD+LSB

PVD method does not utilize the smooth area to hider more secret data, resulting in lower embedding payload. In order to obtain larger payload, Wu et al. developed a method of PVD and LSB to hide data. Their approach is a combination of PVD and LSB. Use a threshold value div to divide the quantization table into two levels, lower level and high level. The threshold value is controlled by user and it is treated as a secret key. Their approach used an idea of PVD where pixel differencing belongs to high level and 3-bits LSB when pixel differencing belongs to low level, respectively.

In their embedding procedure, firstly, a cover image is partitioned into non-overlapping blocks with two consecutive pixels, saying p_x and p_{x+1} . The quantization table, showing in Fig. 1, is used to hide secret data into each block by using hiding strategy of PVD or LSB substitution according block differencing d_i falling into.

Here, the block differencing d_i is computed by $d_i = \lfloor p_x - p_{x+1} \rfloor$. If the block differencing d_i belongs to high level, the embedded method is the same as PVD scheme. In the other words, if the block differencing d_i belongs to lower level, p_x and p_{x+1} are hidden by using 3-bits LSB approach. Let p'_x and p'_{x+1} are the embedded results of p_x and p_{x+1} , respectively. Note, LSB substitution method may cause the new block differencing d'_i falling into high level. Thus, the adjustment strategy is necessary for readjust new block differencing into the same level, lower level. The adjustment strategy is followed as:

$$(p'_i, p'_{i+1}) = \begin{cases} (p'_i - 8, p'_{i+1} + 8), & \text{if } p'_i \geq p'_{i+1} \\ (p'_i + 8, p'_{i+1} - 8), & \text{if } p'_i < p'_{i+1} \end{cases} \quad (2)$$

The secret data are extracted from each blocks is the same order as the embedding phase. The information of quantization table is needed to run the process of data extracting. Moreover, the embedded bits can easily to be extracted according to new block differencing d'_i falling into. Assume that new block differencing d'_i belongs to high level, secret bits is extracted by using the PVD method. On the contrary, new block differencing d'_i belongs to lower level, secret bits is extracted by applying 3-bits LSB method. All the blocks are processed, thus, the secret bits are completely extracted.

← Lower level	High level →			
[0, 16]	[16, 31]	[32, 63]	[64, 127]	[128, 255]

Fig. 1. The quantization table of Wu et al.'s approach

3 Proposed Scheme

In this section, we will introduce our data hiding scheme based on two tiers strategy. Where two tiers strategy represents two quantization tables are used to estimate how many bit will be hidden. We extend the idea of PVD, and combine it with the concept of pixel tolerance, which is defined in [5, 13]. In the PVD of Wu and Tsai, hidden data into each block with two pixels is determined by computing pixel differencing and the optimal range of differencing falling. In their method, two pixels should be tolerated with great changes since the differencing value is larger. Assume that a block has large differencing value; it means that one pixel in a block has smaller pixel value. For example, a block with two pixels is $(P_i, P_{i+1}) = (172, 6)$, and the block differencing d is $166 (= 172 - 6)$. Following up the PVD scheme, value of P_{i+1} is responsible pixel changes for concealing data as the same as value of P_i . It is disobey the concept of pixel tolerance. We therefore propose new technique that applies a threshold TH to evaluate which one pixel should be tolerated with smaller changes. This subsection will elucidate our proposed method in detail, including the embedding and extracting algorithm.

3.1 Data Embedding Algorithm

The proposed method uses the digital images with 256 gray level as cover images. The secret data can be seen as a long bit stream. The number of bits, which is can hidden varies, is decided by the quantization table. The quantization table is designed by user. Here, two quantization tables, Tab_l and Tab_h showing in Tables 1 and 2, are demanded for our proposed. The details of the data embedding algorithm are as follows.

Step 1: Input cover image, and use raster scan manner to divide image into non-overlapping block with two consecutive pixels, called as P_i and P_{i+1} .

Step 2: Check the pixel density and select a quantization table according to following two cases.

Case 1: If the value of two pixels P_i and P_{i+1} are smaller than TH , take a quantization table of Tab_l to run the next steps.

Case 2: If one of two pixels P_i and P_{i+1} or both of them are large than TH , carry a quantization table of Tab_h to run the next steps.

Step 3: Calculate the differencing value d_i between two consecutive pixels in the blocks by $d_i = |P_i - P_{i+1}|$.

Step 4: Find the optimal range R_i from the quantization table according d_i belongs to. Then, obtain the number of embedding k bits and read k bits of secret bit string.

Step 5: Embed k secret bits into P_i and k secret bits into P_{i+1} , respectively, by using LSB substitution. Assume that P'_i and P'_{i+1} be the hidden result of P_i and P_{i+1} , respectively.

Step 6: Apply the modified LSB substitution method to P'_i and P'_{i+1} .

Step 7: Calculate the new differencing value which is defined as:

$$d'_i = |P'_i - P'_{i+1}| \quad (3)$$

Step 8: If d_i and d'_i belong to differencing range, execute the readjust phase as follows.

Case 8.1: $d_i \in R_i$, $d'_i \in R_{i+1}$. If $P'_i \geq P'_{i+1}$, readjust the (P'_i, P'_{i+1}) to be the best choice between $(P'_i - 2^k, P'_{i+1})$, $(P'_i, P'_{i+1} + 2^k)$, and $(P'_i - 2^k, P'_{i+1} + 2^k)$; otherwise, readjust the (P'_i, P'_{i+1}) to be the best choice between $(P'_i, P'_{i+1} - 2^k)$, $(P'_i + 2^k, P'_{i+1})$, and $(P'_i + 2^k, P'_{i+1} - 2^k)$.

Case 8.2: $d_i \in R_i$, $d'_i \in R_{i-1}$. If $P'_i \geq P'_{i+1}$, readjust the (P'_i, P'_{i+1}) to be the best choice between $(P'_i + 2^k, P'_{i+1})$, $(P'_i, P'_{i+1} - 2^k)$, and $(P'_i + 2^k, P'_{i+1} - 2^k)$; otherwise, readjust the (P'_i, P'_{i+1}) to be the best choice between $(P'_i, P'_{i+1} + 2^k)$, $(P'_i - 2^k, P'_{i+1})$, and $(P'_i - 2^k, P'_{i+1} + 2^k)$.

Step 9: Output two pixel values, P'_i and P'_{i+1} , as the stego pixels.

Notably, in Step 8, the best choice, say (b_i, b_{i+1}) , indicates that it satisfies conditions that differencing $d'_i = |b_i - b_{i+1}|$ and d_i belongs the same range and $(b_i, b_{i+1}) \in [0, 255]$, also the value of $|b_i - P_i|^2 + |b_{i+1} - P_{i+1}|^2$ is smaller.

Table 1. The proposed quantization table

Table Name	Ranges	[0, 15]	[15, 63]	[64, $TH-1$]	
Tab _l	Number of embedding bits	2	3	4	
Table Name	Ranges	[0, 15]	[15, 63]	[64, 127]	[128, 255]
Tab _h	Number of embedding bits	4	5	6	7

Table 2. The other proposed quantization table

Table Name	Ranges	[0, 15]	[15, 63]	[64, $TH-1$]	
Tab _l	Number of embedding bits	3	4	5	
Table Name	Ranges	[0, 15]	[15, 63]	[64, 127]	[128, 255]
Tab _h	Number of embedding bits	4	5	6	7

3.2 Data Extracting Algorithm

The process of data extracting is the same as the process of data embedding except the input image is stego image. In addition, the same quantization tables, which are used in data embedding process, are used in here. The data extracting in the stego image is as follows.

- Step 1: Partition the stego image into non-overlapping block of two consecutive pixels using raster scan manner, saying P'_i and P'_{i+1} .
- Step 2: Compute the differencing value d'_i between two consecutive pixels in a block by $d'_i = |P'_i - P'_{i+1}|$, and applies the threshold TH to check the pixel density, following as two cases.
 - Case 1: If the value of two pixels P'_i and P'_{i+1} are smaller than TH , carry a quantization table of Tab_l to run the next steps.
 - Case 2: If one of two pixels P'_i and P'_{i+1} or both of them are not smaller than TH , take a quantization table of Tab_h to run the next steps.
- Step 3: Apply the quantization table and differencing value d'_i to obtain the optimal range R_i which d'_i belongs to.
- Step 4: Get the number of k secret embedding bits according to optimal range R_i , and use k -bits LSB substitution of a pixel to extract k secret bits from P'_i and to extract k secret bits from P'_{i+1} .
- Step 5: Execute Step 2 to Step 4 to extract k secret bits for each block.

After all secret bits have been extracted or all blocks have been processed, the data extracting procedure is done.

4 Experimental Results

In this section, we will demonstrate the experiment of our proposed. In our experimental, we use five gray level images with size 512×512 as cover images, such as Elaine, Lean, Peppers, Sailboat, and Zelda, which are shown in Fig. 2. The secret bit in our algorithm was generated randomly in the program of Visual C++. We employ the

peak-signal-to-noise-ratio (*PSNR*) to estimate image quality between cover image and stego image, which is determined as $PSNR(db) = 10 \times \log_{10}(\frac{255^2}{MSE})$. Where *MSE* represents mean square error. If *PSNR* value is larger than 30 db, the image quality in stego image is accepted for human visual system [13].

Table 3 show the results of our proposed using different quantization tables with $TH = 192$. From this table, the average embedding payloads of our proposed method using different quantization tables are 2.25 bpp and 3.04 bpp (bit per pixel), respectively. The *PSNR* values are above 35 db. So, the stego image quality by using our approach is imperceptible.

We compared the proposed method with other previous approach, such as Wu and Tsai’s PVD method [6], Wu et al.’s PVD and LSB approach [8], and so on. The compared resultant is demonstrated in Tables 4 and 5. According to Table 4, our proposed approach has high performance than previous work in terms of embedding payload and image quality. The Table 5 shows that our performance has higher embedding capacity than simple 3-bits LSB, PVD and LSB, and adaptive LSB method while all the *PSNR* values are more than 36 db.

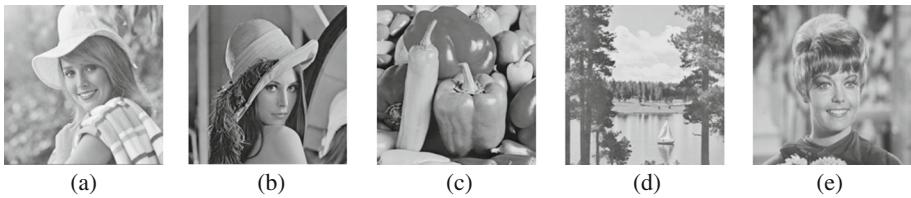


Fig. 2. Five cover images: (a) Elaine; (b) Lena; (c) Peppers; (d) Sailboat; (e) Zelda.

Table 3. The resultant of *PSNR*s and hiding capacity (bits) in our approach using difference tables for five test images.

Images Tables		Elaine	Lena	Peppers	Sailboat	Zelda
Table 1	PSNR	41.044	42.184	42.473	41.082	45.806
	Hidden Bits	624,568	591,214	572,216	629,624	534,666
Table 2	PSNR	34.627	36.268	36.221	35.102	40.139
	Hidden Bits	886,712	853,358	834,360	891,768	796,810

Table 4. Comparisons of Hiding capacities (bits) and average PSNRs between our approach using Table 1 and previous approaches.

	Wu and Tsai's method [6]	Yang et al.'s method [10]	Our approach
Ave. PSNR	41.61	40.35	42.52
Elaine	408,582	419,697	624,568
Lena	406,632	410,854	591,214
Peppers	401,980	408,281	572,216
Sailboat	415,554	430,888	629,624
Zelda	388,584	402,163	534,666

Table 5. Comparisons of Hiding capacities (bits) and average PSNRs between our approach using Table 2 and various approaches.

	Simple 3-bits LSB	Wu et al.'s method [8]	Yang et al.'s method [12] (3-4-5 division)	Our approach
Ave. PSNR	37.90	37.48	38.43	36.82
Elaine	786,432	755,027	816,956	886,712
Lena	786,432	774,970	812,794	853,358
Peppers	786,432	776,160	804,226	834,360
Zelda	786,432	781,306	797,108	796,810

5 Conclusions

In this paper, we proposed an adaptive data embedding scheme to hide secret data into a gray level image. In our scheme, firstly, we divided a cover image into non-overlapping blocks with two consecutive pixels. Then, apply a threshold TH to determine which one pixel could be tolerated with small change, and employ quantization tables to estimate how many bits could embed in two consecutive pixels. We used modified LSB substitution to conceal secret bit into cover image. Moreover, the adjustment phase is used to make the differencing value of two consecutive pixels in the same range, before and after data embedding. The experimental results show that our proposed scheme has good image quality and has better performance than that of previous works.

Acknowledgments. This research was partially supported by the Ministry of Science and Technology of the Republic of China under the Grant MOST 105-2221-E-324-014 and MOST 105-2221-E-153-010.

References

1. Chu, Y.H., Chang, S.: Dynamical cryptography based on synchronized chaotic system. *Inst. Elect. Eng. Electron. Lett.* **35**, 974–975 (1999)
2. Kuo, W.C., Lai, P.Y., Wang, C.C., Wu, L.C.: A formula diamond encoding data hiding scheme. *J. Inform. Hind. Multi. Signal Process.* **6**, 1167–1176 (2015)
3. Anderson, R.R., Peticolas, F.A.P.: On the limits of steganography. *IEEE J. Sel. Areas Commun.* **16**, 474–481 (1998)
4. Chen, C.K., Chen, L.M.: Hiding data in images by simple LSB substitution. *Pattern Recog.* **37**, 469–474 (2004)
5. Yang, C.H.: Inverted pattern approach to improve image quality of the information hiding by LSB substitution. *Pattern Recognit.* **41**, 2674–2683 (2008)
6. Wu, D.C., Tsai, W.H.: A steganographic method for images by pixel-value differencing. *Pattern Recognit. Lett.* **24**, 1613–1626 (2003)
7. Ynag, C.H., Weng, C.Y.: A steganographic method for digital images by multi-pixel differencing. In: *Internatioanl Computer Symposium, Taiwan*, pp. 831–836 (2006)
8. Wu, H.C., Wu, N.I., Tsai, C.S., Hwang, M.S.: Image steganographic scheme based on pixel-value differencing and LSB replacement methods. *IEE Proceed. Vis. Images Signal Process.* **152**, 611–615 (2005)
9. Wang, C.M., Wu, N.I., Tsai, C.S., Hwang, M.S.: A high quality steganography method with pixel-value differencing and modulus function. *J. Syst. Softw.* **81**, 150–158 (2008)
10. Yang, C.H., Weng, C.Y., Tso, H.K., Wang, S.J.: A data hiding scheme using the varieties of pixel-value differencing in multimedia images. *J. Syst. Softw.* **84**, 669–678 (2011)
11. Yang, C.H., Weng, C.Y., Wang, S.J., Sun, H.M.: Varied PVD+LSB evading detection program to spatial domain in data embedding system. *J. Syst. Softw.* **83**, 1635–1643 (2010)
12. Yang, C.H., Weng, C.Y., Wang, S.J., Sun, H.M.: Adaptive data hiding in edge area of images with spatial LSB domain system. *IEEE Tran. Inform. Fore. Secu.* **3**, 488–497 (2008)
13. Wang, S.J.: Steganography of capacity required using modulo operator for embedding secret image. *Appl. Math. Compu.* **164**, 99–116 (2005)
14. Hong, W., Chen, T.S., Luo, C.W.: Data embedding using pixel value differencing and diamond encoding with multiple-base notational system. *J. Syst. Softw.* **85**, 1166–1175 (2012)
15. Mandal, J.K., Das, D.: Steganography using adaptive pixel value differencing of gray images through exclusion of overflow/underflow. In: *International Conference on Computer Science, Engineering and Application, India*, pp. 93–102 (2012)
16. Chen, J.: A PVD-based data hiding method with histogram preserving using pixel pair matching. *J. Image Commun.* **29**, 375–384 (2014)
17. Swain, G.: Adaptive pixel value differencing steganography using both vertical and horizontal edges. *Multimedia Tools Appl.* **75**, 13541–13556 (2016)
18. Hosam, O., Halima, N.B.: Adaptive block-based pixel value differencing steganography. *Sec. Commun. Netw.* **9**, 5036–5050 (2016)

Intelligent Systems

A House Price Prediction for Integrated Web Service System of Taiwan Districts

Chia-Chen Fan^(✉), Shyan-Ming Yuan, Xuebai Zhang,
and Yu-Chuan Lin

Computer Science, National Chiao Tung University,
1001 University Road, Hsinch, Taiwan
wandy260178@yahoo.com.tw, smyuan@gmail.com,
asuracocoa@mail.com, lp01po26399366@hotmail.com

Abstract. Buying a house is not an easy thing for the most people. If you want to buy a house, you must to consider many factors. Such as the house pattern and location. These factors directly or indirectly affect the value of the house value. The current sale of the house only to provide the price and details of house information. There is no provision of the housing prices trend. Hence, this system is a network service for combine the house price forecast and the sale of house information. House buyers make good choice by this house price prediction service. This system use analytical method and forecasting model to forecast house prices. In experimental results, we use hit rate to verification if the forecast interval is reasonable. More than half of six city's hit rate above 75%. It is means our system can help people to buy satisfied house.

1 Introduction

Buying a house is very difficult. If we want to buy a house, we must to consider many factor. For example exterior of house, construction type, pings of building, and age of house etc. Different building affect their value. In addition, it affect house prices that houses in different locations such as different county or district. Different districts will show different price trends in different months. Therefore, buyer must consider the location exterior of house and construction type before they buy the house.

In recent years, there are many scholar studies house prices forecasting. They use a lot of method and algorithm to forecast the house prices trend. Sudipto Guha, Rajeev Rastogi, Kyuseok Shim [2] proposed a hierarchical clustering algorithm. References [1] proposed K-prototype Clustering, Modified K-means Clustering, and Expand K-means Clustering to forecast the house price. M.V. Jagannatha Reddy and B. Kavitha [7] proposed a method to handle numeric attributes and category attributes at the same time. Ming-Yi Shih, Jar-Wen Jheng and Lien-Fu Lai [3] proposed two-stage approach called TMCM algorithm to handle mixed type of attribute information. References [3] use HAC clustering algorithm to group the data. There are many paper studies using different prediction model, references [4] forecast district by linear regression model.

2 System Design and Implementation

2.1 System Structure

Proposed system based on references [5] architecture. We provide buyer an integrated price forecast internet service. Beginning, when the home buyers to enter the system home page, they can choose districts, house parameters and house types. After the house buyer selecting the house parameters, they send the information to the system. The system through references [6] to link the web server. On the other hand, the system pass parameters to server by Python in the background program. Integrated housing price forecasting system architecture is in the Fig. 1:

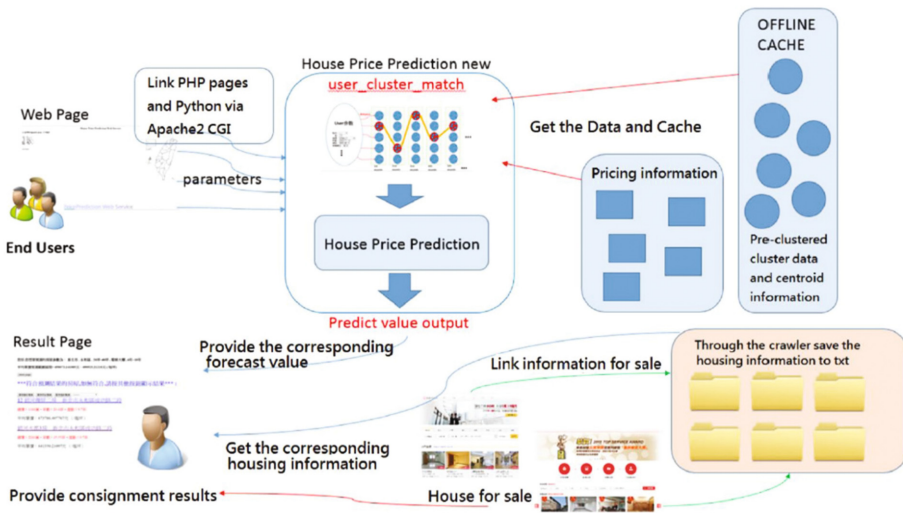


Fig. 1. Integrated housing price forecasting system architecture

2.2 Price Forecast Program

In the following section, we will discuss the price forecast program and how to work the system. This system are two part. There are offline and online stage.

– The offline stage have six parts:

1. Data pretreatment: Remove the outlier, fill the lost value, and standardization of numerical data.
2. Feature engineering: Removal of parking spaces, partitioned the non-urban areas, and compiled the non-urban areas.
3. Cluster grouping: We use the Expand K- means grouping algorithm, distinguish numeric attributes, and category attribute.
4. Cluster pairing: We use the Hungarian pairing algorithm to link months and months, and we will get the corresponding temporal data.

5. Forecast: This system predicts the next month house price use the Brown's double exponential smoothing.
 6. Cluster verification: Find the appropriate parameters for the group number, number of iterations, and time.
- The online stage have three parts to analysis:
1. Find representative data from a monthly cluster: After validation, we can get the appropriate number of clusters, the number of iterations and the number of months training. We use the Expand K-means clustering algorithm to separate each month's data into different clusters.
 2. Comparison of user parameters and clusters: The user chooses the preferred house parameters on the page such as county/city, administrative district, housing type, and house age. We will save these parameters after the send these data to the server. In this step, we will try to find the timing data is closest to the user parameter. The characteristics of the cluster corresponding to each month and the user parameter are the closest. Hence we use the quantification to pairing distance of five representative information per month cluster.
 3. Brown's double exponential smoothing forecasting: This step is housing price forecast. We use the data of house parameters selected by the user. We select the price of the 5 house records closest to the center of the group as the average of the month. And we use this average as the actual value of such housing class into the forecast model to do the forecast. The advantage of this approach is that you can make the program fast. Another advantage is that if the month doesn't meet the user's housing type, we can still use close to the center of the five records as a forecast value. To solve the dilemma when there is no record.

2.3 Percentage Error

We are forecast for house buyers choose the housing parameters in each administrative district. Then we discuss the percentage error between the predictive values by Brown's double exponential smoothing prediction method and the actual value by the actual combination of parameters. We can utilize this data to analyze the reliability of our forecasting model.

3 System Implementation

The system implementation results are two parts. The first part is the forecast interval for all combinations of parameters in each administrative area of the city. And to calculate the hit rate of accord the combination of parameters in the month. Then the first part can verification the forecast interval. The second part is satisfaction survey for the user. This part can verify that our system architecture and design process whether user considers this system convenience or not.

3.1 Experimental Environment and Programming

The system’s experimental environment is to use a Virtual Machine to complete the site deployment and it as a program running the main server. The Virtual Machine’s specifications as Table 1:

Table 1. Integrated housing price forecasting system architecture

OS	CPU	RAM
Ubuntu 14.04 LTS	Intel® Core™ i5-4440 Processor - 3.1 GHz	2 GB

In this system, we analysis the house price forecast by Python as the system’s core. The following Fig. 2 is program architecture.

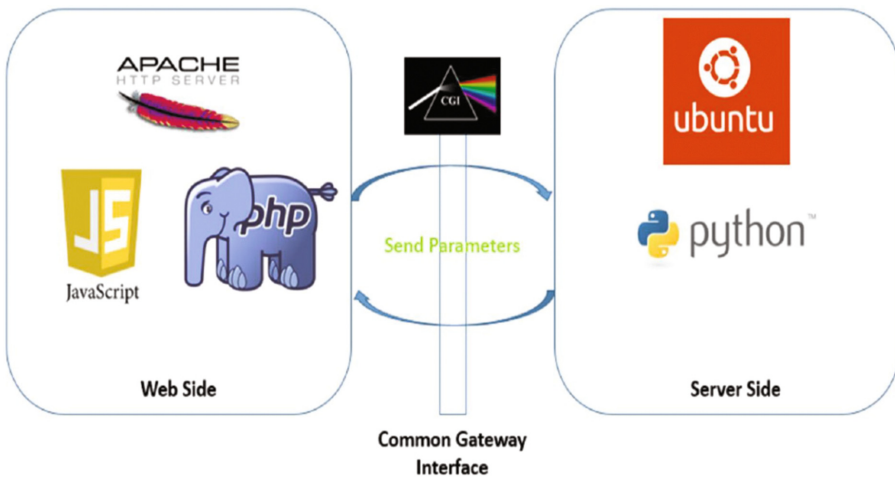


Fig. 2. Program architecture

3.2 Predict Hit Rate

In the first part of the experiment, we make the forecast interval for all combinations of parameters in each administrative area of the city. The forecast interval is 68% predicted interval. Since the predictions result are an interval, not an average value. When we are doing the comparison, we think that each category is basically a trend of a normal distribution. We find the house that matches that type of the month. Then sorted by the price, find out the mean and standard deviation and find the distance between the plus and minus a standard deviation within the house. We compare the range of information and observe if these houses within 68% forecast interval. Final, we calculate the hit rate as following Fig. 3:

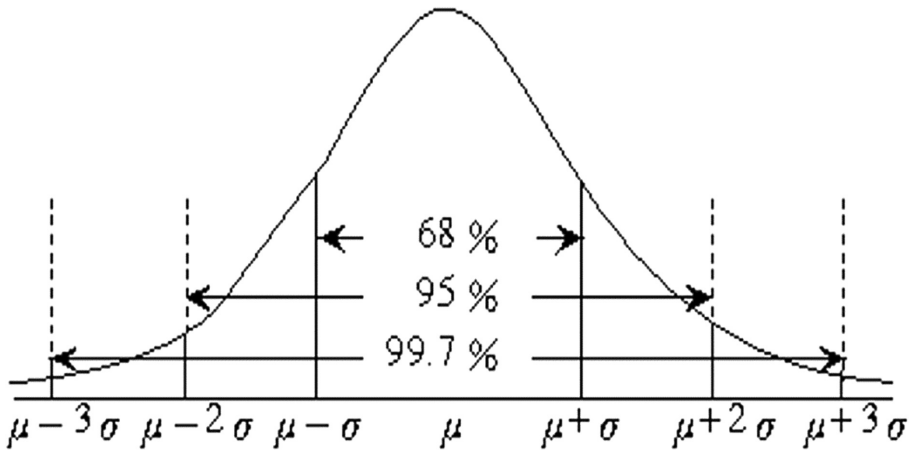


Fig. 3. Forecast interval

There are 120 possible combinations of parameters for each district. For example 4 types of houses, 5 kinds of house size, and 6 kinds of age. We compare and calculate these 120 combinations. Finding the real number of the type of the house in every month. We only consider more than 3 type of the house. Then find these hit rate. We used New Taipei and Kaohsiung city’s house information in January 2015 to forecast April 2015 house situation. The housing forecast hit rate table as following Table 2:

Table 2. House forecast hit rate for new Taipei and Kaohsiung city at April 2015

City	Hit rate				
	100%	75%~100%	50%~75%	25%~50%	0%~25%
New Taipei city	91	29	39	11	43
Kaohsiung city	80	17	28	8	36

From the data to know, New Taipei City and Kaohsiung’s hit rate more than 75% occupied all type of house 56.3% and 57.4%, respectively. The hit rate more than 50% occupied all type of house 74.6% and 74.0%, respectively.

3.3 System Availability Scale

In the second part of the experiment, we design a table to know the availability of this system. And users think this system is good or not. System Availability Scale is a subjective feelings table. The table represents user subjective feelings for this system. We invited 30 volunteer to use this system. And they agree to help us complete this experiment. They are gender, age, basic information, and planning for buying a house as the following Table 3:

Table 3. Users information

Number of people volunteer	Gender		Age		Buy a house Willingness	
	Male	Female	Below 30	30~65	Have	Haven't
30 people	24	6	23	7	6	24

We have 10 question in this system. This table design five level to select. Select one representative very disagree. Select five representative very agree. The system's system Availability Scale as follows Table 4:

The system availability scale score 80.1. This means the house price system is useful for house buyers.

Table 4. System availability scale

	System availability scale	Average	Standard
1.	I will often use this system	3.3	0.8
2.	This system is very complicated	1.6	0.8
3.	This system is easy to use	4.2	0.7
4.	Professionals are required to assist with this system	1.5	1.1
5.	System integration is good	3.8	0.6
6.	This system is too much inconsistent	1.4	0.7
7.	This system is easy to learn	4.3	0.8
8.	This system is cumbersome to use	1.3	0.6
9.	I am very confident in this system	4.0	0.5
10.	Before using this system, I need some knowledge	1.8	1.1

SAS score: 80.1

4 Conclusion

We have proposed an integrated housing price forecast network service. This system provides a simple and easy service. The house buyers don't too much expertise and background for buying a house.

We use the Brown's double exponential smoothing forecast module to design the system. This system is a house price forecast network service and it simple and easy-to-use for house buyers. It can reduce the time for home buyers to inquire about other related sites. Hence, users can save a lot of time to choose a favorite house.

Compliance with Ethical Standards

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Informed consent was obtained from all individual participants included in the study.

References

1. Liang, C.P.: House price prediction system based on open government data. Master thesis, Institute of Network Engineering College of Computer Science, NCTU (2015)
2. Guha, S., Rastogi, R., Shim, K.: ROCK a robust clustering algorithm for categorical attributes. In: 15th International Conference on Data Engineering Proceedings, vol. 25, pp. 345–366. IEEE (1999)
3. Shih, M.Y., Jheng, J.W., Lai, L.F.: A two-step method for clustering mixed categorical and numeric data. *Tamkang J. Sci. Eng.* **13**(1), 11–19 (2010)
4. Chiu, S.C.: Real estate price models of based on real price registration. Master thesis, Institute of Computer Science and Engineering College of Computer Science, NCTU (2014)
5. Apache2. <https://help.ubuntu.com/lts/serverguide/httpd.html>
6. CGI. <http://www.w3.org/CGI/>
7. Reddy, M.V.J., Kavitha, B.: Clustering the mixed numerical and categorical dataset using similarity weight and filter method. *Int. J. Database Theory Appl.* **5**(1), 121–134 (2012)

Commonsense-Knowledge Based Inference Engine

Zhengdao Peng^(✉) and Jean Lai

Hong Kong Baptist University, Kowloon, Hong Kong
16444833@life.hkbu.edu.hk, jeanlai@comp.hkbu.edu.hk

Abstract. Nowadays, more and more industries achieve automatic large-scale production, and also more and more robots undertake the responsibilities of domestic chores. When machines run in industries or robots work in home, they should have abilities to make judgement and abilities to learn from experience, if they want to do their jobs well. In this study, a commonsense knowledge base (CKB) and an inference engine that can support decision making with the commonsense knowledge are built. They can understand human languages, and equip with inferencing abilities, and learning abilities.

Keywords: Inference engine · Commonsense base · Artificial intelligence

1 Literature Reviews

Artificial intelligence is one of the hottest topics of computer science, because AI could help human in many different ways and especially doing what we are not able to do in our lives. Artificial intelligence means systems could simulate such aspects of human intelligence as speech recognition, deduction, inference, creative response, the ability to learn from experience, and the ability to make inferences given incomplete information [1]. One of the fundamental problems, which is encountered when using AI to build universal machines, became known as the general knowledge problem or the commonsense knowledge problem, and while researchers were aware that in an AI system, knowledge would have to be explicitly represented, they did not anticipate the vast amount of implicit knowledge we all share about the world and ourselves [2]. So, we built a system, which is called commonsense-knowledge based inference engine and will be introduced in the following sections, and this inference engine would help AI system to mainly focus on problems in daily lives based on what we called commonsense knowledge base.

Commonsense knowledge is the collection of facts and information that an ordinary person is expected to know, and the commonsense knowledge problem is the ongoing project in the field of knowledge representation (a sub-field of artificial intelligence) to create a commonsense knowledge base: a database containing all the general knowledge that most people possess, represented in a way that it is available to artificial intelligence programs that use natural language or make inferences about the ordinary world [3]. Nowadays, many different kinds of commonsense knowledge bases are existed and there are two of the most frequently-used commonsense knowledge bases

which are called Cyc knowledge base and ConceptNet. Cyc is an AI project which tried to construct a comprehensive knowledge base of everyday commonsense knowledge base. Cyc was launched in 1984 and is as part of Microelectronics and Computer Technology Corporation, and CycL is Cyc representation language, which is based on predicate calculus [4]. ConceptNet is a semantic network based on the Open Mind Common Sense (OMCS), which was launched in 1999 at the MIT Media Lab, and is the network includes nodes and edges, which respectively represent concepts and assertions of commonsense about concepts. Concepts represent sets of closely related natural language phrases, which could be noun phrases, verb phrases, adjective phrases, or clauses [5]. We seldom see people applying Cyc knowledge base and ConceptNet in real life applications due to the limited usefulness of the outcomes generated. In this study, we tried to construct a new commonsense knowledge base and an inference engine which aim to be used by real-life applications.

2 System Development

In this section, the implementation process and structure of the inference engine and commonsense knowledge base is covered.

2.1 Overview of the Inference Engine

Initially, we aim to design a system for better decision making with commonsense knowledge.

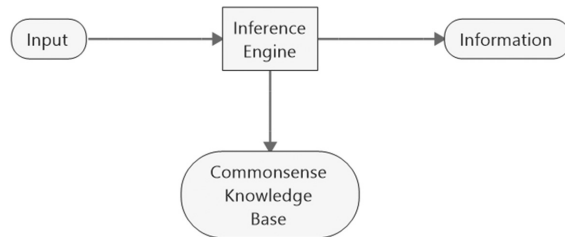


Fig. 1. System overview

From Fig. 1, the workflow of the system is described. Firstly, users input a query into the system, and then system gets data from users' input and commonsense knowledge base (CKB), and run programs inside the inference engine. Finally, users would get information, which could help users to make decisions, from the system. For example, user wants to know whether he or she should go out today or not, and inputs this question into the system. After system works for a few seconds, system would tell you that 'it's rainy outside, so it's not a good idea to go out today.' And after getting this information, the user will decide to stay at home based on the information.

2.2 System Architecture

In this system, there are two major parts. One is inference engine, which mainly deal with data that it gets from outside, and gets results by analyzing the data. Another part is CKB, which mainly provides inference engine with data or information, and which consist of a large number of commonsense knowledges.

Architecture of the CKB. We built a CKB as simple as possible. CKB is divided into many indexes based on the domain of commonsense knowledges, and each index contains a large number of knowledges. So far, in our prototype, this CKB has five indexes, and they are: Time Index, Verb Index, Question Words Index, Command Index and Response Index. The number of indexes could be increased based on needs. Each index has two columns. One column is consisted of knowledges in human language, and another column is composed of data or information, which system could understand. The index (like a mapping table) helps to map human language into machine understandable form. Fig. 2 obviously shows the process of how commonsense knowledge works in each index. On the left side of the index, facts are the data that system actually knows before, and then facts would compare with each tuple of knowledge in the index, and the final results would be sent to the inference engine.



Fig. 2. Indexes in CKB

Table 1. Examples of time index

Input	Output
today	0
yesterday	-1
tomorrow	1
monday	Mon
mon	Mon
tuesday	Tue
tue	Tue
sunday	Sun
sun	Sun
weekday	[Mon, Tue, Wed, Thu, Fri]
weekend	[Sat, Sun]

For example, we have an index, called Time Index, and it seems like Table 1, and we assume that the input fact is ‘weekend’. Then, the fact would compare with the data in Input column of each tuples of knowledge, and finally, inference engine would get the result, ‘[Sat, Sun]’, which inference engine could understand. Other indexes would also work in the same way.

Architecture of the Inference Engine. In Fig. 3, the detailed process of how the system works is included. User inputs a sentence into the system. A Sentence Parser would parse the input, save it in an array form, and then sent the array to the inference engine. The array is then passed to the Inference Trigger and return a new array (Fig. 4). And the last process is that intermediate results would be passed to the Response Trigger and the system would return the final answer. Sentence Parser, Inference Trigger and Response Trigger would get commonsense knowledge from the CKB.

In addition, a learning process, in which user can provide feedback to the system, is included. The feedback would impact the results of the Inference Trigger. This design is to enhance the inferencing result progressively. The details about each of three processes and the learning process are as follows.

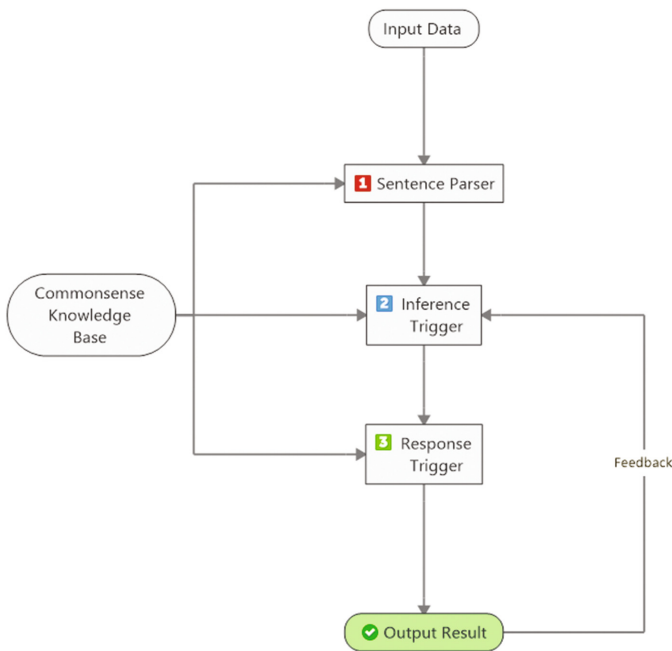


Fig. 3. The working process of inference engine

First Process. Figure 4 shows the details of first process. Sentence Parser would get a query from users, and this query would be sent to Transformer, which would get each word from the query and transform each word into lowercase. All these words would be grouped as an array. Each item of the array would compare with each tuple of Time Index, Verb Index and Question Words Index respectively, and after comparing, all results would be grouped into a new array. For example, user input a query, ‘Should we go out today?’, into the system, and then Transformer would first catch the query and produce an array, like ‘[should, I, go, out, today]’. Then, each item of this array would be compared with each tuple of knowledges in Time Index, Verb Index and Question

Words Index respectively, so finally the output for next step is a new array, like '[should, go_out, 0]', which inference engine could understand.

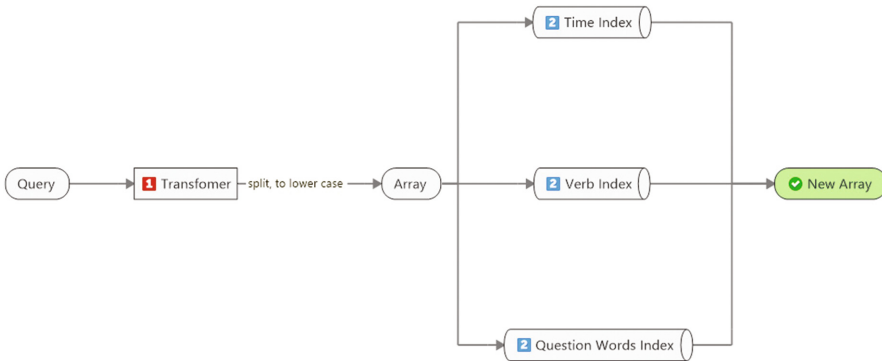


Fig. 4. Detailed process of sentence parser

Second Process. Figure 5 indicates the detail of second process. In second process, the array, which is the result of the first process, is compared with knowledge in Command Index, and then, Inference Trigger would get a set of commands and corresponding ratios. Ratio, is a number or null, and could be calculated by the system in the Learning Process. Each of commands and ratios, and the array which is from the first process compose a new array, so Inference Trigger would get a set of new arrays. For each array, the corresponding function would be triggered after we get the array, and also, we would get the corresponding ratio and result. For example, the result of the first array is '[should, 0, go_out]', and after sending it into Command Index, we could get [should, 0, goout_weather, 3] and [should, 0, goout_schedule, 5]. Then, goout_weather function and goout_schedule function would be triggered, so after all functions finish their operations, output of this process could be [{should, 0, goout_weather, 3, result 1}, {should, 0, goout_schedule, 5, result 2}].

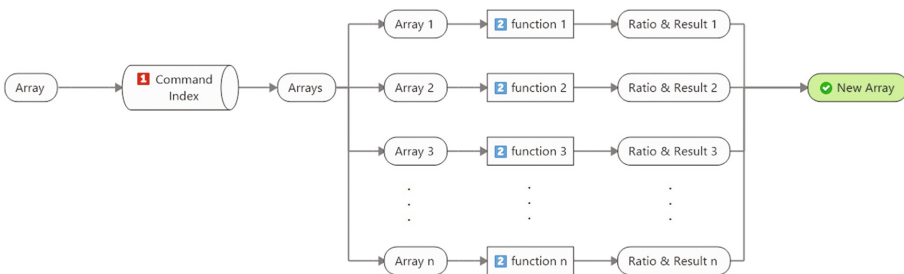


Fig. 5. Detailed process of inference trigger

Third Process. In third process, the result from the second part is as the input of the third process. Firstly each item of the input array compare with each tuple of knowledges in the Response Index, and system gets the new array like, $[\{\text{should}, 0, \text{goout_weather}, 3, \text{response 1}\}, \{\text{should}, 0, \text{goout_schedule}, 5, \text{response 2}\}]$, and then it is passed to the Map Function (see Fig. 6). Each item of new array would be passed into all constraints in the Map Function, items, which fulfill all constraints, would be grouped as a new array. Then, Map Function compares each pair of items in the array and use Roulette Selection to choose only one response to be outputted. Roulette Selection would randomly choose one result based ratios of all items, which means those items, which have higher ratio, have higher probabilities to be outputted as final result.

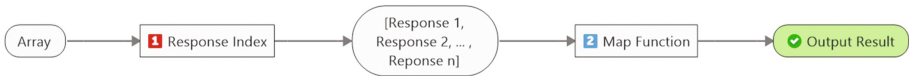


Fig. 6. Detailed process of response trigger

Learning Process. Figure 7 shows that the learning process of this system. In this process, the main idea is that the feedback from the result, which was outputted to users, changes the ratio of the corresponding command in the second process, and as we mentioned before, the ratio would also have influence on the probability of the response appearing as the final result. The probability of the system asking for feedback is depending on the original ratio. If the ratio is null, which means it have no feedback before, then it would have a higher probability to ask for feedback. If the ratio is not null, then it would have a lower chance to ask for feedback.

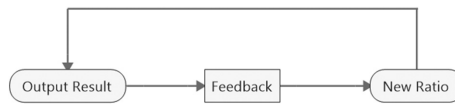


Fig. 7. Detailed process of learning process

3 System Modeling

3.1 The CKB

At the very beginning of the process of modeling, we thought it would be great to combine a new engine with a commonsense knowledge base which was already existed, so a freely-used commonsense-knowledge base, which called ConceptNet, was chose for this project. ConceptNet is a semantic network, which was initially designed for computers to understand human using words. Then before building the model, it is necessary to test this commonsense knowledge base by using it with a simple application. Firstly, we downloaded the database from official website and modified the structure of the knowledge base, like deleting the data of other languages except

English from the knowledge base. This knowledge base would be helpful to get some results which are related to what we input, but we are not so sure about if all results from this knowledge base would have strong relation to our input. And then, all modified data was automatically inputted into MongoDB and used some codes to apply this commonsense knowledge base. But finally, the results contained many words, which are quite unrelated to what we input, for example, we would get ‘sea’, ‘sluice’ and ‘change’ from inputting ‘go out’. So, because of the unreasonable results, we decided to build a new commonsense knowledge base for this system. In this CKB, we built some indexes, like Time Index, Verb Index and Command Index, and stored the whole CKB into MongoDB, and different index would be used in different part of inference engine. Also, some knowledge was inserted into indexes, so each of indexes would provide inference engine with those knowledge, which could act as rule for inference engine.

3.2 The Inference Engine

After building the CKB, the inference engine was begun to be developed. The program language is Node.js. Firstly, each of Sentence Parser, Inference Trigger and Response Trigger was respectively developed, and each of them should be correctly connected to the corresponding index in the CKB. And then, these three parts was combined as a whole part. And the final process is to test if this inference engine could run as expected.

3.3 The Web Application

After inference engine is developed, it would be necessary to build a web application to show how inference engine work well, and this application could be an assistant, like Siri or Cortana in some extent, so a webpage was built to show a chat room, which was connected to inference engine. After building the webpage, these two parts are still disconnected, so socket would be used to send and receive data from one to another one. In the chat room, user could input their queries and inference would give answers to users. Personal user data from the webpage would be directly stored into MongoDB.

4 Conclusion

Users concluded that this design can solve the limitations in existing CKB and inference engine. This kind of system is combining inference engine with commonsense knowledge base, and it could be the brain of the artificial intelligence, which could help a robot to have abilities of inference and abilities of learning. So, when we build a huge and structured commonsense knowledge base, artificial intelligence could understand a large number of human languages after used by inference engine. This system could work well in all circumstance, but it still need to be improved to some extent, and it could be more automated and structured in the future.

References

1. Microsoft.: Microsoft Computer Dictionary. 5th edn. Microsoft Press, Washington (2002)
2. Page in PSY371. <http://psych.utoronto.ca/users/reingold/courses/ai/>. Accessed 30 Apr 2017
3. The Cognitive Computing Era: Commonsense Knowledge in BPM. <https://bpm.com/bpm-today/blogs/1007-the-cognitive-computing-era-commonsense-knowledge/>. Accessed 30 Apr 2017
4. Stork. David, G.: HAL's Legacy: 2001's Computer as Dream and Reality. MIT Press, Cambridge (1997)
5. Havasi, C., Speer, R., Alonso, J.: ConceptNet: a lexical resource for common sense knowledge. In: Recent Advances in Natural Languages Processing V: Selected Papers from RANLP 2007, pp. 269–279. John Benjamins Pub. Co, Philadelphia (2007)

Analysis of Users' Emotions Through Physiology

Bohdan Myroniv, Cheng-Wei Wu^(✉), Yi Ren, and Yu-Chee Tseng

Department of Computer Science, National Chiao Tung University,
Hsinchu, Taiwan, ROC
cww0403@nctu.edu.tw

Abstract. Most of the existing studies focus on physical activities recognition, such as running, cycling, swimming, etc. But what affects our health, it is not only physical activities, it is also emotional states that we experience throughout the day. These emotional states build our behavior and affect our physical health significantly. Therefore, emotion recognition draws more and more attention of researchers in recent years. In this paper, we propose a system that uses off-the-shelf wearable sensors, including heart rate, galvanic skin response, and body temperature sensors to read physiological signals from the users and applies machine learning techniques to recognize their emotional states. We consider three types of emotional states and conduct experiments on real-life scenarios with ten users. Experimental results show that the proposed system achieves high recognition accuracy.

Keywords: Emotion recognition · Machine learning · Physiological signals · Wearable devices

1 Introduction and Background

Emotion is an integral part of our health and has a huge impact on our bodies, mental health and behavior. Poor emotional health can weaken our immune system, making us more likely to get colds and other infections. According to American Psychological Association (APA), [10] more than half (53%) of Americans report personal health problems as a source of stress. Stress that has left unchecked can contribute many health problems, such as high blood pressure, heart disease, obesity and diabetes. According to a study conducted by the American College Health Association (ACHA) [11], a considerable proportion of students said that mental health problems affected their academic performance, causing them to dropping courses, or receiving low grades in the classes.

Most of the existing studies of emotion recognition are based on *visual camera-based approaches* [2, 3, 6] and *audio speech-based approaches* [1, 4, 5, 7]. The visual camera-based approaches [2, 3, 6] use image processing technologies to detect users' emotions. However, they require the users to be well seen by the cameras. Besides, their accuracy may be degraded in a poorly lit environment. The main idea of audio speech-based approaches is to use the speakers' loudness, tone and intonation patterns in a speech to detect their emotions. However, conversational properties may vary in different cultures and nationalities, which may lead such system to suffer from

worse recognition results. Less attention was paid to physiological signals analysis for emotion recognition using wearable devices. However, detecting users' emotions using physiological signals allows collecting reliable data because the autonomous nervous system cannot be controlled consciously by users themselves. Such data will not affect by factors, such as light requirements in video-based approaches or cultural peculiarities in audio-based approaches.

In this paper, we apply machine learning techniques to recognize users' emotions using low-cost wearable devices. We consider three types of sensors in wearables: heart rate sensor, temperature sensor and galvanic skin response (abbr. GSR) sensor, to collect sensor signals related to users' emotions. Then, we apply sliding window-based segmentation method to segment collected data and extract candidate features from the segments. After that, we feed extracted features to classifiers to identify the type of users' emotional states. We implement the prototype of the proposed methods on Arduino platform and evaluate the effectiveness of the proposed methods on real-life scenarios with ten users. Experimental results show that the proposed system is very effective and achieves high recognition accuracy.

2 The Proposed Method

In this section, we introduce the proposed methods and explain the procedure of physiological signals processing for obtaining user's emotional states. Fig. 1 shows the three types of non-invasive biosensors used in this work: heart rate sensor, body temperature sensor and galvanic skin response sensor. The three sensors are connected to Arduino Yun microcontroller to collect users' physiological signal data. Data are collected with 10 Hz sampling rates. The collected sensor data are sent to a server via Bluetooth for further processing. Figure 2 shows the Input-Process-Output diagram of the proposed system. We consider recognition of three types of emotional states: positive, neutral and negative. Received signals go through four processing phases: pre-processing, segmentation, feature extraction and emotion classification.

Pre-processing Phase. During the data collection, we observe that the collected data tend to have missing values during a short period (i.e., one to two seconds) just after the connection between the devices is established. But afterwards, no such problem happens. Therefore, we cut off the first two seconds' data from the collected data to avoid the interference of the missing data. Such issue may happen due to the ongoing establishment of the Bluetooth connection.



(a) Heart rate sensor.



(b) Temperature sensor.



(c) GSR sensor.

Fig. 1. The sensors used in this work

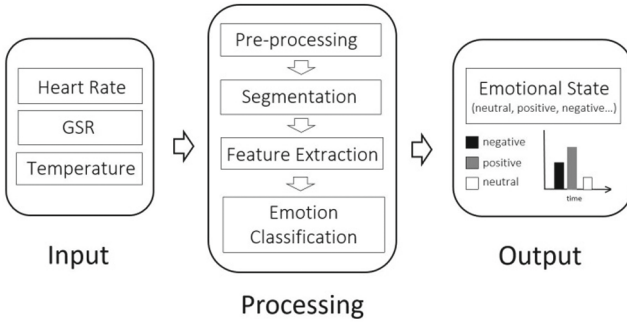


Fig. 2. The input-process-output diagram of the proposed system

Segmentation Phase. In this phase, we applied overlapped sliding window-based segmentation to segment the collected sensor data. The window size for segmentation is set to θ ($\theta > 0$) seconds and the overlapped window is set to δ ($0 \leq \delta \leq \theta$) seconds. In the experiments, we will evaluate the effect of varied window sizes and overlapped sizes on the recognition result of the proposed system (Fig. 3).

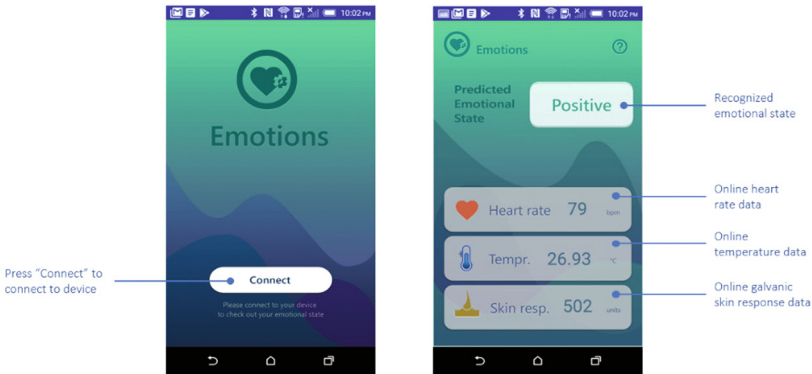


Fig. 3. The prototype result of the implemented App

Feature Extraction Phase. In this phase, we extract candidate features from the sensor signals in each segment. Let $A = \langle a_1, a_2, \dots, a_n \rangle$ be a time series data of a sensor in a segment, where a_i ($1 \leq i \leq n$) is the i -th sample in A . The extracted candidate features are shown in Table 1.

Emotion Classification Phase. After extracting candidate features, we then use the feature data to build classifiers using classification algorithms in Weka [12]. We consider the following six different types of classification algorithms: K-Nearest Neighbor (abbr. KNN), J48 Decision Tree (abbr. J48), Naïve Bayes (abbr. NB),

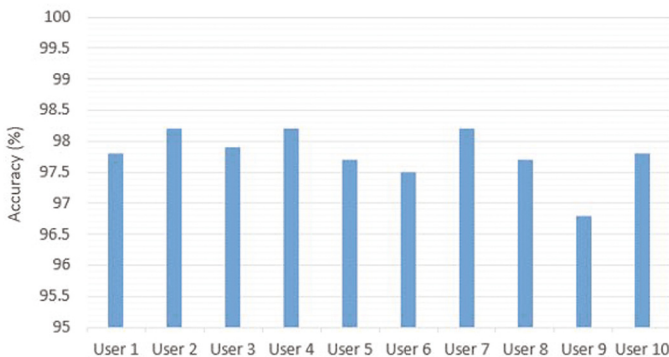
Table 1. The extracted candidate features

Feature	Formula
Mean	$\frac{1}{n} \sum_{i=1}^n a_i$
Variance	$\frac{1}{n} \sum_{i=1}^n (a_i - \mu(A))^2$
Energy	$\frac{1}{n} \sum_{i=1}^n (a_i)^2$
Average absolute difference	$\frac{1}{n-1} \sum_{i=2}^n a_{i-1} - a_i $
Average absolute value	$\frac{1}{n} \sum_{i=1}^n a_i $
Skewness	$\frac{\frac{1}{n} \sum_{i=1}^n (a_i - \mu(A))^3}{\left[\frac{1}{n} \sum_{i=1}^n (a_i - \mu(A))^2\right]^{3/2}}$
Kurtosis	$\frac{\frac{1}{n} \sum_{i=1}^n (a_i - \mu(A))^4}{\left[\frac{1}{n} \sum_{i=1}^n (a_i - \mu(A))^2\right]^2}$
Zero crossing rate	$\frac{1}{2} \sum_{i=2}^n \text{sign}(a_i) - \text{sign}(a_{i-1}) $
Mean Crossing Rate	$\frac{\sum_{i=2}^n \text{sign}(a_i - \mu(A)) - \text{sign}(a_{i-1} - \mu(A)) }{2}$
Root mean square	$\sqrt{\frac{1}{n} \sum_{i=1}^n (a_i)^2}$

Random Tree (abbr. RT), Support Vector Machine (abbr. SVM) and Multilayer Perceptron Neural Networks (abbr. MP).

3 Prototype Implementation

In this section, we introduce the prototype result of the proposed system. We developed an Android application to recognize user's emotional state online. It consists of Arduino side and Android side. On the Arduino side, heart rate, body temperature and GSR sensors will continuously collect user's physiology signals and send them to the smartphone via Bluetooth. On the Android side, the collected data will go through pre-processing, segmentation, feature extraction and emotion classification phases. Figure 4 shows the prototype of the implemented App. On the first screen of the app user clicks "Connect" button to connect the App with the Arduino device.

**Fig. 4.** The accuracy of each personalized model

When Bluetooth connection is established, application will go to second screen where user can see three boxes on which shown his/her detected heart beats per minute (show. as *Heart rate* in the interface), GSR (abbr. as *Skin resp.* in the interface) and temperature (abbr. as *Temp.* in the interface) data from the interface in a real-time fashion. Detected user's emotional state is shown in the upper right side of the app. Online recognition is done by using the pre-trained classifier inside the smartphone.

4 Experimental Evaluation

We recruited ten participants in our experiments, where there are seven males and three females. Their ages are between twenty-two and thirty. All the experiments are conducted in separate room. We also asked participants to leave their phones out during the experiments. Emotion triggering is a very important part of the experiment. We use two affective emotion photo databases in our experiments. The first one is Geneva Affective Picture Database (GAPED) [8] provided by Swiss Center for Affective Sciences, and the second one is International Affective Picture System (IAPS) [9] provided by Center for The Study of Emotion and Attention. We choose twenty photos for each type of emotional states (i.e., positive, neutral, and negative) from the above mentioned databases. Each photo was shown for five seconds. We have tried different orders of data collection, i.e. different emotion stimuli were shown in different days, and different consistency. Such approach was applied to minimize dependencies of one emotional condition from another one.

We held our experiments three times. The data collected from the first experiment was never used in the processing. This is because participants themselves left comments that they felt nervous and strange due to the bunch of sensors and cables connected to their arms, which would influence participants' emotion and lead to noise data. Therefore, the data collected from the first experiment is not used. However after the first experiment, our participants were more familiar with the sensors setup and it wasn't disturbing for them to wear these sensors in the next data collection rounds. Therefore, the data collected from the second and third experiments are used in the processing.

We test the recognition performance of the proposed sensing system using different types of classifiers, including KNN, J48, NB, RT, SVM, and MP. We apply 10 fold cross-validation to do the performance evaluation. We consider two kinds of models called *group model* and *personalized model*. In group model, the emotion classifier is constructed by considering all the participants' physiology signals in the training phase. While in personalized model, the classifier is constructed by considering individual participant's physiology signals in the training phase.

Table 2 shows the accuracy of the group models under the different size of segmentation. In this experiment, the overlapping size δ is set to 0 s. In Table 2, we can see that the J48 classifier performs better than the other five classifiers. Besides, when the window size θ is set to 1 s, the J48 classifier achieves 90.35% recognition accuracy. In Table 2, the SVM and NB classifiers are not very effective, their average accuracies are 45.59% and 39.91%, respectively.

Table 2. The accuracy of the group models under the different size of segments.

	1 s	2 s	3 s	4 s	5 s
RT	83.69	82.90	82.58	73.29	80.52
J48	90.35	84.49	86.48	87.25	80.55
NB	39.46	39.16	48.34	41.62	34.00
KNN	72.76	74.75	76.57	80.45	79.56
SVM	47.11	42.94	42.64	44.84	49.00
MP	86.28	82.30	79.87	77.28	72.00

Then, we test the effect of δ on the overall accuracy of the proposed system. In this experiment, the window size θ is set to 5 s. Table 3 shows the accuracy of the group models when the overlapping size is varied from 0.5 to 4.5 s. As shown in Table 3, overlapping size is set to 4.5 s, the KNN classifier has the best recognition accuracy, which is up to 97.31%.

Table 3. The accuracy of the group models under different sizes of overlapping.

	RT	J48	NB	KNN	SVM	MP
0.5 s	84.68	81.92	39.63	78.37	49.54	77.92
1.0 s	79.51	84.33	38.95	79.11	42.57	80.72
1.5 s	85.31	86.01	38.11	77.62	44.40	76.92
2.0 s	85.58	85.88	44.44	84.08	48.34	83.18
2.5 s	86.53	90.77	44.38	85.28	46.38	83.54
3.0 s	87.41	89.27	41.38	84.58	45.28	85.15
3.5 s	89.82	88.62	49.11	87.27	49.85	88.47
4.0 s	93.11	92.31	46.30	92.11	48.41	88.92
4.5 s	94.22	95.46	52.14	97.31	48.41	91.78

In the next experiment, we fix the parameters θ and δ to 5 and 4.5 s, respectively, and evaluate the precision, recall and F-Measure of the proposed system using KNN classifier for each class. Table 4 shows the experimental results. As shown in Table 4, the system using KNN classifier has good recognition result for each class.

Table 4. The precision, recall and F-Measure for the group model using the KNN classifier.

Class	Training data			Testing data		
	Precision	Recall	F-Measure	Precision	Recall	F-Measure
Positive	98.4	98.0	98.2	97.3	97.4	97.4
Neutral	97.4	97.9	97.6	98.7	97.6	98.1
Negative	98.5	98.3	98.4	98.7	97.6	98.1

Next, we train model for each participate to test the recognition accuracy for personalized models. Because there are ten participants, ten personalized models were built. In this experiment, we set the parameters θ and δ to 5 and 4.5 s, respectively, and consider the KNN classifier. Figure 4 shows the classification results for each personalized model of each user. As shown in Fig. 4, the average accuracy of all the personalized models is 97.78%. Moreover, the maximum and minimum accuracies are 98.2% and 96.8%, respectively.

5 Conclusion

In this work, we consider the heart rate, body temperature and galvanic skin response sensors to design a wearable sensing system for effective recognition of user's emotional states. The proposed system allows recognizing three types of emotions, including positive, neutral and negative, in an online fashion. We apply the machine learning technology to process the physiology signals collected from the user. The process consists of four main phases: data pre-processing, data segmentation, feature extraction, and emotion classification. We implement the prototype of the system on Arduino platform with Android smartphone. Extensive experiments on real-life scenarios show that the proposed system achieves up to 97% recognition accuracy when it adopts the k -nearest neighbor classifier. In the future work, we will consider more types of user's emotional states and consider to find a correlation between these emotions and physical activities for more diverse and novel applications.

Acknowledgement. This work was support in part by the Ministry of Science and Technology of Taiwan, ROC., under grant MOST 104-2221-E-009-113-MY3, 105-2221-E-009-101-MY3, 105-2218-E-009-004.

References

1. Amin, S., Andriluka, M., Bulling, A., Müller, M.P., Verma, P.: Emotion recognition from embedded bodily expressions and speech during dyadic interactions. In: IEEE International Conference on Affective Computing and Intelligent Interaction, pp. 663–669 (2015)
2. Bulut, M., Busso, C., Deng, Z., Kazemzadeh, A., Lee, C.M., Lee, S., Narayanan, S., Neumann, U., Yildirim, S.: Analysis of emotion recognition using facial expressions, speech and multimodal information. In: International Conference on Multimodal Interfaces, pp. 205–211 (2004)
3. Bilakhia, S., Cowie, R., Eyben, F., Jiang, B., Pantic, M., Schnieder, S., Schuller, B., Smith, K., Valstar, M.: The continuous audio/visual emotion and depression recognition challenge. In: International Workshop on Audio/Visual Emotion Challenge, pp. 3–10 (2013)
4. Aswathi, E., Deepa, T.M., Rajan, S., Shameema, C.P., Sinith, M.S.: Emotion recognition from audio signals using support vector machine. In: IEEE Recent Advances in Intelligent Computational Systems, pp. 139–144 (2015)
5. Dai, K., Fell, J.H., MacAuslan, J.: Recognizing emotion in speech using neural networks. In: International Conference on Telehealth/Assistive Technologies, pp. 31–36 (2008)

6. Chakraborty, A., Chakraborty, U.K., Chatterjee, A., Konar, A.: Emotion recognition from facial expressions and its control using fuzzy logic. *IEEE Trans. Syst. Man Cybern. Part A: Syst. Humans* **39**(4), 726–743 (2009)
7. Pao, T.L., Tsai, Y.W., Yeh, J.H.: Recognition and analysis of emotion transition in Mandarin speech signal. In: *IEEE International Conference on Systems Man and Cybernetics*, pp. 3326–3332 (2010)
8. Dan-Glauser, E.S., Scherer, K.R.: The Geneva affective picture database (GAPED): a new 730-picture database focusing on valence and normative significance. *Behav. Res. Methods* **43**, 468–477 (2011)
9. Bradley, M.M., Cuthbert, B.N., Lang, P.J.: International Affective Picture System (IAPS): technical manual and affective ratings. In: *NIMH Center for the Study of Emotion and Attention, University of Florida* (1997). <http://csea.php.ufl.edu/media.html>
10. American Psychological Association. <http://www.apa.org/index.aspx>
11. Consequences of Poor Mental Health. www.campushealthandsafety.org/mentalhealth/consequences/
12. Weka 3: Data Mining Software in Java. <http://www.cs.waikato.ac.nz/ml/weka/>

Research on Temperature Rising Prediction of Distribution Transformer by Artificial Neural Networks

Wenxin Zhang, Jeng-Shyang Pan, and Yen-Ming Tseng^(✉)

Fujian Provincial Key Laboratory of Data Mining and Applications/School of Information Science and Engineering, Fujian University of Technology, Fuzhou 350108, China

2465212498@qq.com, jengshyangpan@gmail.com, swk1200@qq.com,

Abstract. In order to predict the temperature rising of the distribution transformer by applying the artificial neural networks (ANNs) method analyze experimental data with the actual measured data and compared with the actual measured value to reach the relative errors investigation. The historical data of the working day are divided into three periods according to the varying loadings trend of load change emotion as the peak period, the general time period and the valley period. In experimental results, The average relative error of the peak period is 2.05%, the average relative error of the general period is 1.69%, the average relative error of the valley period is 1.25 %, and the working day average relative error is 1.60% for a day 24 hours. By Ann's derivation the result has a very good prediction rate at temperature rising of distribution transformer.

1 Introduction

Electricity power brings people a lot of convenience as that human liberation from a lot of constraints and improves people living levels and quality in recent society. Of course, that is by invention and contribution the transformer. Safe and stable economic operation of distribution system must be relay the safe and stable operation of the transformer but the temperature transformer of coils directly determines the life of the transformer and the stability of the operation of electricity supply. Accurate calculation and prediction of the transformer internal temperature rise is the transformer research and operation and maintenance of the key. From view of the engineering practice, there are many factors that affect the temperature rise of the transformer and some factors are even difficult to determine. By analyzing the various factors related to the transformer temperature rise to make sure the main factors to from the ANNs input variables for train and recall data and to establishment of the corresponding mathematical model by artificial neural networks (ANNs). Furthermore, it is predict the change of the temperature rise of the transformer and compare between both of the actual measured value and the predicted value to find the relative error. There are many monitoring methods to monitoring the temperature rising of transformer and the purpose of monitoring is to find the influence factors of temperature rise from the historical data and to preceding

the uninterrupted supply the power for users. Usage for many forecasting trend has many methods to support such as wavelet network monitoring [1], rule base of expert system [2], model of ARIMA for load forecasting [3], time series [4], gene method [5], and the use of neural network method [6–8] and so on. The factors affecting the temperature rise factors of the transformer there are many studies have pointed out that such as transformer hysteresis loss [10], harmonic distortion [9]. Because of the good adaptability of the neural networks which are applied to the load forecasting [11–13] posses well approached function. Also use the fuzzy theory [14, 15] to apply the load forecasting. In this paper, the neural network is predicted the temperature rising of the transformer.

2 Artificial Neural Networks (ANNs) Method

Artificial neural network (ANN) is widely connected by large number of non-linear processing units and posses with large-scale parallel processing of information, self-learning ability and high fault tolerance that is likewise the brain structure of the simplified, abstract and simulation. Among the many kind of the ANN model that the most widely methods usage is the forward multilayer back-propagation (BP) model that contain input layer, the hidden layer, and the output layer. Every layer consists of by many neurons or called node by full connection with the nearby nodes of layer and the hidden layer not only by one there are can be more. Figure 1 is neural network structure of back-propagation and contain two layers of hidden layers.

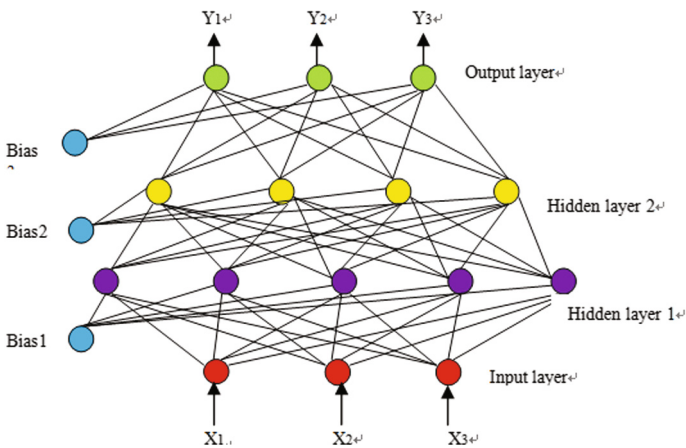


Fig. 1. Back-propagation structure of neural network

The architecture with multi-layer neural network of ANN has the ability of high nonlinear mapping and symmetry discrimination which through the repeated training by a series of training points or called training set to map the nonlinear relationship with the input variables and the respect output. The algorithm consists of neural

network forward transmission and error back propagation and which in forward calculation stage and the input information is processed from the input layer through the hidden layer and passed to the output layer in order. In the network training stage, the forward propagation cannot reach the desired output and then transferred to the reverse propagation which base on the error information between with the input and output difference to modifying the neurons of the weights and thresholds by every neuron connecting with the nearby neuron of different layer to minimum the output error to reach the respect output value and approach the accurate classification that is belong to the supervision of learning network in ANNs. The training data is to use a known input variable and a known output variable to train a convergent neural network and till the response data uses a known input variable to predict an unknown output variable.

3 Selection of Key Variables

Transformer in operation status which parts of coils temperature rising due to core hysteresis loss, eddy current loss and coil copper loss by loadings. Which temperature will higher than the room temperature that produces the thermal flow in the area by the form of radiation, conduction, etc., that the heat generating and dissipation to reach equilibrium will made the temperature of each part tends to in stable. Iron loss (hysteresis loss and eddy current loss) is a fundamental loss which is closely related to the transformer structure and the rated voltage and almost no change. But it cannot be reduced or eliminated during operation is the copper loss (line loss) that varies with the load and significant. Figure 2 is shown the daily temperature rising and loading curve of transformer. In this figure we can find that are in midnight and mooring 0: 00 to 8: 00 A.M., its called valley time interval, the loading is very small and the temperature of transformer is the downward trend. From 8: 00 to 8: 30 A.M., its called load increasing speedup time interval, the loading quickly increasing and temperature of transformer is quickly increasing, too. From 8: 00 A.M. to 16: 00 P.M., it is called peak load time interval, which the temperature slight increases or decrease following the loading of transformer when the load increases decrease. From 18: 00 to 19: 00 P.M., its called load decrease falling down time interval, the loading quickly decrease and temperature of transformer is quickly decreased, too. And from 19:30 to 23: 30 P.M., it's called valley time interval and in this period the loading is very small but in 19:30 to 20:00 P.M. the temperature of transformer is increasing and in 20:30 to 23.30 P.M. the temperature of transformer is decrease which is related to reactive power, and reactive power with voltage harmonic distortion is inversely proportional to the relationship.

Figure 2 is the tread curve of temperature following the reactive power that said the relationship is chaotic. Figure 3 is the tread curve of temperature following the reactive power that said the relationship is chaotic. Figures 4 and 5 are weekly temperature of transformer V.S. voltage harmonic distortion curve and temperature of transformer V.S. current harmonic distortions, respectively.

The internal temperature of the power transformer and the contact temperature are not exactly the same but the both relationship can be expressed by nonlinear function. If the power transformer for the self-cooling, Temperature is closely related with the indoor temperature and dissipation rate.

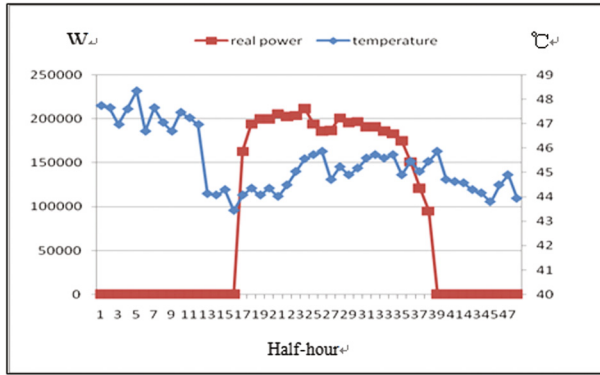


Fig. 2. Daily temperature and load curve

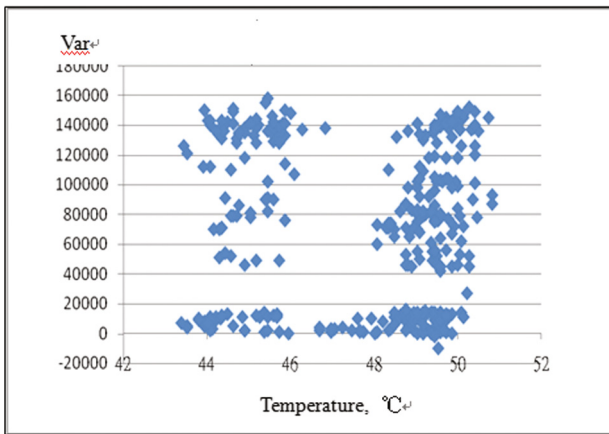


Fig. 3. The tread curve of temperature following the reactive power

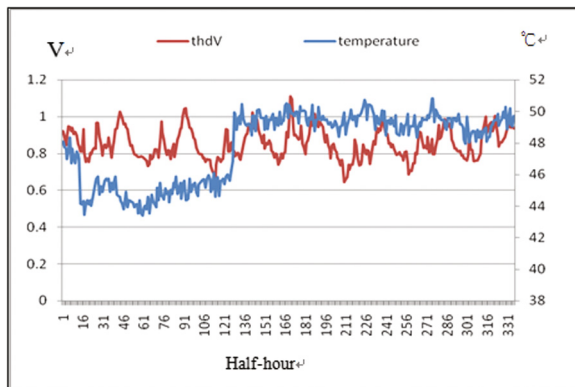


Fig. 4. Weekly temperature of transformer V.S. voltage harmonic distortion curve

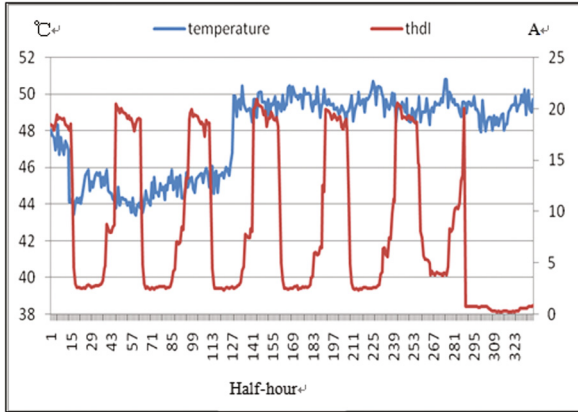


Fig. 5. Weekly temperature of transformer V.S. current harmonic distortion curve

It selects the six variables with relatively high correlation for temperature of transformer to carry out the experiment and forecasting the rising trend of temperature. There are six input variables of the neural network: real power P, reactive power Q, total voltage harmonic distortion (thdV), total Current harmonic distortion (thdI), indoor temperature (idT) and last half hour indoor temperature((lhhT) to investigate the temperature of transformer for output.

4 Classification of Prediction and Experimental Results

4.1 Classification for ANN

Figure 6 is zone of classification of prediction temperature of transformer which consists of five zones. Compare with the daily load curve that classify for three ANNs those are (1) zone I and V are called valley time interval, (2) zone III is called load increasing speedup time interval, and (3) zone II and IV are called load increasing speedup time interval and load decrease time interval. According to the time series by

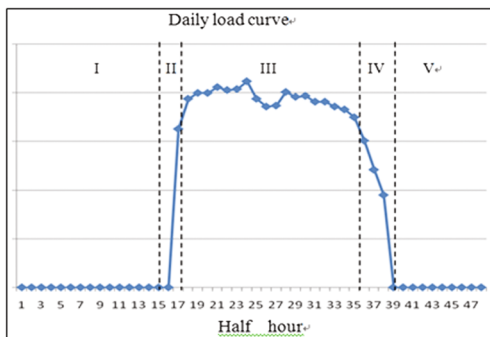


Fig. 6. Zone of classification of prediction temperature of transformer by ANN

apply the upper ANNs to forecasting the very half hour temperature of transformer and by combination the time series of forecasting temperature can form the daily temperature curve of transformer.

4.2 Experimental Results

From a number of actual measurement data selected from the previous week’s working day and to get the training set, furthermore, feed into the ANNs to train the networks till converged. Furthermore, it is from the second week of Tuesday’s measured data to form the recall set and to feed into the converged network to get the temperature forecasting of transformer. According to the actual training and prediction results, there are chosen from the training methods of MATLAB class of neural network with a small percentage of the relative error of the training method which can not only effectively achieve the expected results, save time and reduce the error value.

Table 1. Percentage of relative error in valley time interval

Time series (half-hour)	Actual temperature value (P.U.)	Forecasting temperature value (P.U.)	Relative error (%)
0	0.8296	0.8197	1.19
0.5	0.8228	0.8263	0.43
1	0.8333	0.8211	1.46
1.5	0.8328	0.8179	1.79
2	0.8233	0.8244	0.13
2.5	0.8313	0.8262	0.61
3	0.8202	0.8149	0.65
3.5	0.8228	0.8127	1.23
4	0.8256	0.8173	1.01
4.5	0.8303	0.8142	1.94
5	0.8195	0.8098	1.18
5.5	0.8235	0.8095	1.70
6	0.8183	0.8134	0.60
6.5	0.8178	0.8032	1.79
7	0.819	0.801	2.20
7.5	0.8148	0.8049	1.22
19.5	0.829	0.8155	1.63
20	0.8326	0.8145	2.17
20.5	0.8273	0.8165	1.31
21	0.8323	0.8169	1.85
21.5	0.8225	0.8142	1.01
22	0.8237	0.817	0.81
22.5	0.8125	0.8236	1.37
23	0.8301	0.8204	1.17
23.5	0.8282	0.825	0.39

Average relative error: 1.23%

Table 1. Percentage of relative error in valley time interval and the average relative error is 1.23%. Table 2. Percentage of relative error in load increasing speedup time interval and load decrease time interval and average relative error: 1.69%. Table 3. Percentage of relative error in load increasing speedup time interval and Average relative error: 1.23%. And the daily average relative error is 1.60%. Figure 7 is the daily relative error curve of distribution transformer.

Table 2. Percentage of relative error in load increasing speedup time interval and load decrease time interval

Time series (half-hour)	Actual temperature value (P.U.)	Forecasting temperature value (P.U.)	Relative error (%)
8	0.8085	0.8104	0.24
8.5	0.8152	0.8426	3.36
18	0.8286	0.8498	2.56
18.5	0.8333	0.8228	1.26
19	0.8189	0.8275	1.05
Average relative error: 1.69%			

Table 3. Percentage of relative error in valley time interval

Time series (half-hour)	Actual temperature value (P.U.)	Forecasting temperature value (P.U.)	Relative error (%)
9	0.8246	0.8215	0.38
9.5	0.8101	0.8274	2.14
10	0.8136	0.8224	1.08
10.5	0.8200	0.8327	1.08
11	0.8246	0.8385	1.69
11.5	0.8106	0.843	4.00
12	0.8131	0.8363	2.85
12.5	0.8233	0.8245	0.15
13	0.8173	0.8361	2.30
13.5	0.8225	0.8341	1.41
14	0.8288	0.8404	1.40
14.5	0.8252	0.845	2.39
15	0.8333	0.856	2.72
15.5	0.8329	0.8595	3.19
16	0.8244	0.8543	3.63
16.5	0.831	0.8456	1.76
17	0.8313	0.8479	2.00
17.5	0.8308	0.8376	0.82
Average relative error: 2.05%			

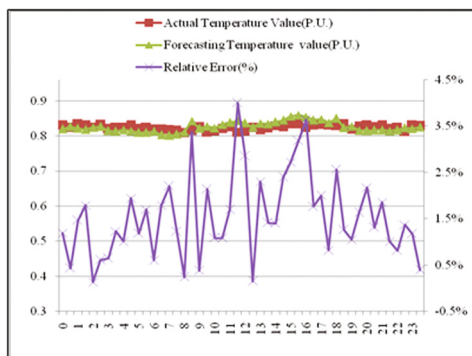


Fig. 7. Daily relative error curve of distribution transformer

5 Conclusion

According to the artificial neural network theory and training method by Traincgp in the BP network model are used to predict the core or coils temperature rising of distribution transformer. This kind of prediction method has the advantages of fastness and accuracy compared with the traditional numerical calculation method because the neural network has the capability of parallel computing, distributed information storage, fault tolerance and strong self-learning function that made forecasting system can be used convenience in the application to improve prediction accuracy. The working day average relative error of temperature rising forecasting is 1.60% for 24 hours a day. Furthermore, the method has high application value and broad application prospect in the performance analysis and optimization design of the transformer.

References

1. Huang, Y.-C., Huang, C.-M.: Evolving wavelet networks for power transformer condition monitoring. *IEEE Trans. Power Deliv.* **17**(2), 412–416 (2002)
2. Wang, Z.Y., Liu, Y.L., Griffin, P.J.: A combined ANN and expert system tool for transformer fault diagnosis. *IEEE Trans. Power Deliv.* **13**(4), 1224–1229 (1998)
3. Contreras, J., Espinola, R., Nogales, F.J., Conejo, A.J.: ARIMA models to predict next-day electricity prices. *IEEE Trans. Power Syst.* **18**(3), 1014–1020 (2003)
4. Nogales, F.J., Contreras, J., Conejo, A.J., Espinola, R.: Forecasting next-day electricity prices by time series models. *IEEE Trans. Power Syst.* **17**(2), 342–348 (2002)
5. Ling, H., Leung, F.H.F., Lam, H.K., Lee, Y.S., Tam, P.K.S.: A novel genetic-algorithm-based neural network for short-term load forecasting. *IEEE Trans. Ind. Electron.* **50**(4), 793–799 (2003)
6. Barghinia, S., Ansarimehr, P., Habibi, H., Vafadar, N.: Short term load forecasting of Iran national power system using artificial neural network. In: *Proceedings of the IEEE Power Tech*, vol. 3 (2001)
7. Niu, D.X., Wang, H.Q., Gu, Z.H.: Short-term load forecasting using general regression neural network. *IEEE Conf. Mach. Learn. Cybern.* **7**, 4076–4082 (2005)

8. Senjyu, T., Takara, H., Uezato, K., Funabashi, T.: One-hour-ahead load forecasting using neural network. *IEEE Trans. Power Syst.* **17**, 113–118 (2002)
9. Kenedy, S.P., Ivey, C.L.: Application, design and rating of transformers containing harmonic currents. In: *Conference Record of 1990 Annual Pulp and Paper Industry Technical Conference IEEE* (1990)
10. de Le'on, F., Semlyen, A.: A simple representation of dynamic hysteresis losses in power transformers. *IEEE Trans. Power Deliv.* **10**(1) (1995)
11. Mori, H., Ogasawara, T.: A recurrent neural network for short-term load forecasting. In: *Proceedings of the Second International Forum on Applications of Neural Networks to Power Systems*, pp. 395–400 (1993)
12. Taylor, J.W., Buizza, R.: Neural network load forecasting with weather ensemble predictions? *IEEE Trans. Power Syst.* **17**(3), 626–632 (2002)
13. Iizaka, T., Matsui, T., Fukuyama, Y.: Novel daily peak load forecasting method using analyzable structured neural network. In: *Transmission and Distribution Conference and Exhibition 2002, Asia Pacific IEEE/PES*, vol 1, pp. 394–399 (2002)
14. Abu-EI-Magd, M.A., Findlay, R.D.: New Approach Using Artificial Neural Network and Time Series Models for Short Term Load Forecasting Neural Networks, *IEEE CCECE 2003, Canadian Conference*, vol. 3, pp. 1723–1726 (2003)
15. Wang, Z.Y., Liu, Y.L., Griffin, P.J.: A combined ANN and expert system tool for transformer fault diagnosis. *IEEE Trans. Power Deliv.* **13**(4), 1224–1229 (1998)

Development of Audio and Visual Attention Assessment System in Combination with Brain Wave Instrument: Apply to Children with Attention Deficit Hyperactivity Disorder

Chin-Ling Chen^{1,2(✉)}, Yung-Wen Tang³, Yong-Feng Zhou¹,
and Yue-Xun Chen⁴

¹ Department of Computer Science and Information Engineering,
Chaoyang University of Technology, Taichung 41349, Taiwan, R.O.C.
clc@mail.cyut.edu.tw, sia4412@yahoo.com.tw

² School of Information Engineering, Changchun University of Technology,
Changchun 130600, Jilin, China

³ School of Physical Therapy, Chun Shan Medical University,
Taichung 40201, Taiwan
tangyw@csmu.edu.tw

⁴ Department of Information Management, Chaoyang University of Technology,
Taichung 41349, Taiwan, R.O.C.
kl247811@yahoo.com.tw

Abstract. The main purpose of this study mainly for children with attention deficit hyperactivity disorder (ADHD) to attention test, combined the development of the Visual and Audio Attention Test System (VAAT). We use Conners Kiddie Continuous Performance Test Second Edition (K-CPT 2) and VAAT testing 16 children with ADHD between 4 to 7 years old and compare the results to observe the K-CPT 2 and VAAT assessment at the same time. We observe the differences in visual and audio of VAAT when they do wrong or do right response via the brain wave variability. The experimental results show that most of the children have inattention in K-CPT 2. The K-CPT 2 and VAAT raw scores at the same time efficient of the response time, standard deviation and change rates are significant. It appears positive correlation. Our system provides good assessment for children with ADHD.

Keywords: Brainwave · K-CPT2 · ADHD · Assessment

1 Introduction

Attention deficit hyperactivity disorder (ADHD) is a common nerve stunt in children. The global prevalence rate of ADHD children is 5.3% [1]. There are three types of ADHD: inattention, impulse and combined type (contains both of the above) [2]. To assess ADHD is a series of complex task. It requires parents, teachers and clinical interviews, observation, psychological assessment and nerve growth check comprehensive investigation [3].

In assessing the ability of children attention most use Continuous Performance Test (CPT) to detect, common testing tools including Test of Variables of Attention (TOVA) [4], Integrated Visual and Auditory CPT (IVA-2) [10] and Conners CPT 3 (CPT 3) [5], Conners Kiddie Continuous Performance Test 2nd Edition (K-CPT 2) [6] etc. These assessment software used primarily to assist the doctors to be a diagnose ADHD's reference tools. The participants are requested to select a specific and respond the stimulus during the CPT assessment. There are many types of stimuli include graphics, letters, numbers and sound [7]. Many studies supported the practicality of the CPT for children with ADHD. It is more efficient in omission error and commission error between normal children and children with ADHD [7].

In addition, more studies are increasing by EEG or CPT to diagnose the children with ADHD [7]. In many literatures, they indicate the ADHD is a brain disease, especially in front of the area leaves and cortex, which is responsible for maintaining attention and control behavior, it can identify these exceptions and convenient diagnostic instruments study [7]. Measure the brain waves is a non-invasive and can provide many of the wave-band, the following are the different brain wave band and features [8], as shown in Table 1.

Table 1. Brain wave frequency and characteristics

Band	Frequency (Hz)	Characteristics
Delta wave	0.1 Hz–3 Hz	Deep sleep, unconscious state
Theta wave	4 Hz–7.5 Hz	Intuition, light sleep
Alpha wave	8–12 Hz	Relaxed without drowsiness, calm
Beta wave	13–30 Hz	Low frequency: concentrate and relax High frequency: excitement, anxiety
Gamma wave	30–50 Hz	Deep focus

In order to assess children with ADHD, this study developed a visual and audio attention test system to analyze the attention of children's related issues. This study used brain wave instrument to measure the current situation of children and collect the brain waves changes immediately of the assessment process.

The purpose of this study was to compare the K-CPT 2 and Visual and Audio Attention Test System (VAAT) at the same time efficient and observe the visual and audio variability. Analyze the brain wave changes when the children with ADHD done the VAAT.

2 Method

2.1 K-CPT 2

This study used K-CPT 2 to evaluate the children's attention related issues between 4 to 7 years old. Test rules will request the participants to respond to the target property and do not respond to non-target on computer screen's picture (commonly appears in

the daily lives of children, such as football picture). There are 5 blocks in the assessment. A block contains 30 targets and 10 non-target property. The test time is 7.5 min. And the picture interval is 1.5 and 3 s. Each picture display time is 0.5 s.

K-CPT 2 in the assessment of the subject's concentration problem which is divided into four degrees. K-CPT 2 includes inattentiveness, impulsivity, sustained attention and vigilance issues. 4 directions will give a diagnosis from mild to severe symptoms. Various degrees include a number of parameters. The parameters are described below:

- (1) Detectability (DPR): Identify the target and non-target of capacity.
- (2) Omissions (OMI): The proportion of the target property missing.
- (3) Commissions (COM): Non-target property appears to be incorrect response ratio.
- (4) Perseverations (PRS): Response time is less than 0.1 s.
- (5) Hit Reaction Time (HRT): The target appears to press the key in time of HRT.
- (6) HRT Standard Deviation (HRTSD): The coherence of the response speed.
- (7) Variability (VAR): The response is consistent variability.
- (8) HRT Block Change (BLKCH): The reaction speed changes of each block.
- (9) HRT Inter-Stimulus Interval (ISICH): The reaction speed changes of different speed block.

2.2 MindWave

In this study, we used the Mindwave brain cubic of Neurosky to develop the application of science and technology [9]. The Mindwave brain cubic uses a single point dry-electrode sensor on the left front of the prefrontal (FP1) of brain waves, referred to points in the left ear lobe (A1), compared to the level of medical equipment to conduct a compact and inexpensive implementation. It can receive the signal through a complicated process and output EEG brain electrical power spectrum, including High Alpha wave, Low Alpha wave, High Beta wave, Low Beta wave, Mid gamma wave and Low gamma wave, Delta wave, and Theta wave and the eSense index of Neurosky provided parameter; including attention and relaxation, etc. It transfers these data to computer per second via blue tooth mechanism.

2.3 Unity 3D

The assessment software VAAT of this study which is designed by Unity 3D game engine and the programming languages are C# and JAVA.2.4. System architecture and processes. VAAT is based on the evaluation algorithm and the computation of brain wave value. These assessment algorithms including identification rate, omission rate, error rate, impulse rate, reaction time, reaction standard deviation, change rate, reaction block change, response capability of speed change and record the brain wave value of the assessment.

VAAT begins with inputting basic personal information, select visual and audio attention test. Before the test, it will allow participants to take the exercise. Let the participants fully understand the assessment content and next step. The test process will record the participant's answer; include wrong and right response and response time. Once the test is terminated, the record of the participants contains identification rate,

omission rate, error rate, impulse rate, reaction time, reaction standard deviation, change rate, reaction block change, speed change response capability and average brain wave (high/low Alpha, high/low Beta, Delta, Theta, mid/low Gamma wave).

2.4 Test Design and Comparison

The study is designed mainly with cartoon animals black and white picture as the main direction. The following is the target and non-target property content and presentation:

1. Assessment time and stimulate property appears: The assessment time is 8.5 min. The way of the stimulate property appears has two-way: 1.5 and 3 s per time interval. After appearing 20 stimulate properties, the interval will change once. The appearing time of the stimulate property is 0.5 or 0.75 s. It was designed into five phases.
2. Target and non-target property content: Target property is commonly seen pictures of animals (such as a dog, cattle and sheep, etc.) and 20 animals' sounds, non-target property is cat picture and cat sounds. Table 2 shows the comparison of our designed VAAT, CPT 3, K-CPT 2 and IVA-2.

Table 2. Comparison of the assessment system of K-CPT 2, CPT 3, IVA-2 and VAAT

	K-CPT 2	CPT 3	IVA-2	VAAT
The stimulus display	500 ms	250 ms	500 ms	500–750 ms
Stimulus appears	Picture	Picture	Picture and sounds	Picture and sounds
Stimulus interval	1.5 s/3 s, two changes	1 s/2 s, 4 s, three changes	1.5 s	1.5 s/3 s, two changes
Press	target	target	target	target
Non-target (quantity/display mode)	1/ball	1/X	2/number 2, sound 2	2/cat pictures, cat sound
Target (quantity/display mode)	10/life common tool	25/except X	2/number 1, sound 1	20/except cat and cat sound
Target and non - target ratio	15:5	16:4	42:8/8:42	18:2
Age	4 to 7	Over 8	Over 6	4 to 7
Test time	7.5 min	14 min	15 min	8.5 min
block/subblock/trail	5/2/20	6/3/20	5/2/50	5/2/20

The differences between VAAT and K-CPT 2 are that we involved audio stimuli, visual and audio stimuli interact with each other. It can directly realize the differences of visual and auditory. In contrast to IVA-2, the presence of stimulants in IVA-2 was fixed in 1.5 s. The appearing stimulus of our system was divided into 1.5 and 3 s. In the design of the picture using cute black and white animal patterns can make children easy to understand the assessment content.

Before starting the assessment, we will explain the rules. The rules are: either seeing the cat's picture or hear the cat's sound, the participant cannot press the button; otherwise, appear the other animal's picture or sound should press the button fast. When press the "evaluation start" key, it will start immediately. Figure 1 shows the assessment screen, the top right shows the current user's brain waves focused value, the lower right of the assessment is shown for the remaining time. If it affects the participant response, it can be closed when press the "fork" button.



Fig. 1. Assessment screen

2.5 Introduce the VAAT Parameters

The evaluation parameters of the VAAT include omission rate, error rate, reaction time, impulse rate, etc. The following are the VAAT parameters:

- (1) Identification rate: the ratio of the number of objects to the number of non-target objects is the ability to distinguish the identification.
- (2) Omission rate: the target no hit reaction and all the need to press the button ratio for the missed rate.
- (3) Error rate: When the non-target appears, without the reaction, the actual error response and all non-target ratio of the error rate.
- (4) Impulse rate: When the target appears, need to press the button reaction, the reaction time is less than 0.1 s ratio.
- (5) Reaction time: When the target appears, need to press the reaction button, the target appears to the actual button reaction time called reaction time.
- (6) Reaction standard deviation: the standard deviation of all the reaction time as the reaction time standard deviation.
- (7) Change Rate: Calculate the standard deviation of the response time of each block, subtract the first block from the last block as the reaction variability.
- (8) Reaction block change: the test time is divided into six blocks to calculate the response time of each block, and the difference between the first block and the difference of the first block is the change of the reaction zone.
- (9) Speed change response capability: different stimulus interval time to calculate the individual average reaction time changes, the slowest interval of the average reaction time minus the fastest time interval of the average reaction time difference for the rate of change response capacity.

Because the above nine parameters contain visual and auditory parameters, so we divided them into nine visual and auditory parameters. The total is 27 parameters.

2.6 Experimental Tools

The experimental hardware devices are: Mindwave Mobile, notebook computer and software applications: combined the development of the Visual and Audio Attention Test System (VAAT), K-CPT 2, SPSS ver.20. The questionnaire paper uses the parental version of SNAP-IV (ADHD Screening).

2.7 Experimental Objects and Processes

We found 16 subjects aged 4 to 7 years old. They had a ADHD in the doctor's diagnosis and made a SNAP-IV (ADHD screening) questionnaire, 4 children with inattention type, 3 had impulse type and 9 merge type.

The process begins with a K-CPT 2 test, and after the test, the VAAT test is performed 10 min after rest. At the beginning of the VAAT, the participant wears the brain wave instrument before the assessment. When the assessment is completed immediately show the value of the assessment after completing the assessment and record the brain waves. The test site is mainly in the small classroom of the hospital, the environment is bright and quiet.

2.8 Statistical Analysis

We analyzed the correlation between K-CPT 2 and the original fraction of VAAT parameters using SPSS statistical software analysis. The correlation degree was expressed as r , $r = 0.1-0.39$ is low correlation, $r = 0.4-0.69$ is moderate correlation, $r = 0.7-0.99$ is highly correlated. The pairwise sample T test is used to analyze the mean difference between the two sets of continuous variables. The first one we used to analyze the difference between visual and auditory, the second part analyze the differences between brain waves in the two groups when the non-target occurs. Statistical analysis was significantly defined as $p < 0.05$ and more significantly is $p < 0.01$.

3 Result

The K-CPT 2 test results showed that most of the participants had a problem of inattention (from light to strong), and one third of the participants had problems with vigilance and sustained attention, but the impulsivity type and the combined in impulsivity problem are not found.

3.1 VAAT and K-CPT 2 Concurrent Validity

Table 3 shows the correlation analysis of SPSS statistical software with the original scores of VAAT and K-CPT 2 parameters of 16 participants. The results showed that the reaction time ($r = 0.515$, $p < 0.05$), the change rate ($r = 0.672$, $p < 0.05$) belongs

to moderate positive correlation. In addition, the reaction standard deviation ($r = 0.727$, $p < 0.01$) is highly positive correlation. The visual-part of identification rate ($r = -0.423^*$, $p < 0.05$) belongs to moderate negative correlation, and the reaction time ($r = 0.538$, $p < 0.05$) belongs to moderate positive correlation, and reaction standard deviation ($r = 0.713$, $p < 0.01$) are highly positive correlations. The audio-part of reaction standard deviation (0.620 , $p < 0.05$) belongs to moderate positive correlation.

Table 3. VAAT and K-CPT 2 concurrent validity

VAAT - K-CPT 2	Overall parameter correlation	Visual part correlation	Auditory part correlation
Identification rate - DPR	0.448	-0.423*	-0.307
Omission rate - OMI	0.172	0.224	0.048
Error rate - COM	0.312	0.289	0.316
Impulse rate - PRS	0.180	0.249	-0.067
Reaction time - HRT	0.515*	0.538*	0.426
Reaction standard deviation - HRTSD	0.727**	0.713**	0.620*
Change Rate - VAR	0.672**	0.417	0.492
Reaction block change - BLKCH	-0.203	-0.304	-0.024
Speed change response capability - ISICH	-0.143	-0.045	-0.329

$r = 0.1-0.39$ is low correlation, $r = 0.4-0.69$ is moderate correlation, $r = 0.7-0.99$ is highly correlated, * $p < 0.05$, ** $p < 0.01$

There is no correlation mostly between VAAT and K-CPT 2 correlation analysis. There may be existed different test way between VAAT and K-CPT 2. It may be join the hearing stimulus such that the participants were different cognitive. Therefore, the correlation of VAAT and K-CPT 2 parameters are not very high.

3.2 Brainwave Analysis

In this study, each participant in the test of the non-target objects appear, there is the test of the current two seconds before the average of the brain waves and the wrong test of the current two seconds before the average of the average. The analysis method was compared using the paired sample T test of the statistical software SPSS. In Table 4, we can find that the difference between the two waves of the first two seconds before the occurrence of the non-target object is that the average Low Beta wave and the Delta wave of the answer are significantly lower in the test. The wrong two seconds before the subject's attention decreased, showing a vent state.

Table 4. Non-target object when the answer and the wrong two seconds before the brain wave difference

	Do right to brainwave mean (standard deviation)	Do wrong to brainwave mean (standard deviation)	T value	Correlation/ Significantly
High Alpha	35512 (25070)	32646 (24007)	1.157	0.265
Low Alpha	45013 (25670)	42697 (19340)	0.774	0.451
High Beta	27214 (184612)	26160 (16513)	0.573	0.575
Low Beta	30626 (31056)	25992 (24760)	2.384	0.031*
Mid Gamma	11445 (6202)	10741 (6130)	0.666	0.516
Low Gamma	21755 (13288)	19890 (12493)	0.910	0.377
Theta	174382 (88226)	149439 (70973)	2.064	0.057
Delta	627702 (190878)	528588 (212374)	2.370	0.032*

*p < 0.05

4 Conclusions

The development of this study using the combination of brainwave instrument development audiovisual focus assessment of children with ADHD, from the results can be found, VAAT and K-CPT 2 parameters of the correlation is not high. It may be due to different ways of testing, or the number of participants is small. The results of the parameters are not high in the relevance. In the future, visual and auditory is separated test, perhaps visual test and K-CPT 2 at the same time will be high degree of validity. In addition, the differences between VAAT and K-CPT 2, VAAT can be compared participants in the visual and auditory attention to what kind of differences. In the VAAT test, a simple animal picture and animal sounds were used to make the child easy to understand the assessment content to make the appropriate response.

In the analysis of the participants of the brain waves, the statistical analysis shows that when the target appears, to do the wrong time compares to do the right previous second the low Beta wave and the Delta wave down case, it presents that the participants cannot concentrate on something and brain empty before 2 s. The designed system offers a very good assistance reference for doctor diagnosis the children with ADHD.

Acknowledgments. This research was supported by the National Science Council, Taiwan, R.O.C., under contract number MOST 103-2632-E-324-001-MY3, MOST 106-2221-E-324-013 and MOST 106-2622-E-305-001-CC2.

Compliance with ethical standards

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Informed consent was obtained from all individual participants included in the study.

References

1. Rietz, E.D., Cheung, C.H.M., McLoughlin, G., Brandeis, D., Banaschewski, T., Asherson, P., Kuntsi, J.: Self-report of ADHD shows limited agreement with objective markers of persistence and remittance. *J. Psychiatr. Res.* **82**, 91–99 (2016)
2. Hayward, A., Tomlinson, A., Neill, J.C.: Low attentive and high impulsive rats: a translational animal model of ADHD and disorders of attention and impulse control. *Pharmacol. Ther.* **158**, 41–51 (2016)
3. Berger, I., Cassuto, H.: The effect of environmental distractors incorporation into a CPT on sustained attention and ADHD diagnosis among adolescents. *J. Neurosci. Methods* **222**, 62–68 (2014)
4. Llorente, A.M., Voigt, R., Jensen, C.L., Fraley, J.K., Heird, W.C., Rennie, K.M.: The test of variables of attention (TOVA): internal consistency (Q1 vs. Q2 and Q3 vs. Q4) in children with attention deficit/hyperactivity disorder (ADHD). *Child Neuropsychol.* **14**(4), 314–322 (2008)
5. Conners, C.K.: *Conners Continuous Performance Test Third Edition (CPT 3)*. Multi-Health Systems Inc., U.S.A. (2014)
6. Conners, C.K.: *Conners Kiddie Continuous Performance Test Third Edition (Conners K-CPT2)*. Multi-Health Systems Inc., U.S.A. (2015)
7. Kim, J.W., Lee, Y.S., Han, D.H., Min, K.J., Kim, D.H., Lee, C.W.: The utility of quantitative electroencephalography and integrated visual and auditory continuous performance test as auxiliary tools for the attention deficit hyperactivity disorder diagnosis. *Clin. Neurophysiol.* **126**, 532–540 (2015)
8. Katona, J., Farkas, I., Ujbanyi, T., Dukan, P., Kovari, A.: Evaluation of the NeuroSky MindFlex EEG headset brain waves data. In: 2014 IEEE 12th International Symposium on Applied Machine Intelligence and Informatics (SAMII), January 2014
9. NeuroSky Inc. <http://neurosky.com/>

Decision Support Systems

Markov Queuing Theory Approach to Internet of Things Reliability

Thi Thi Zin^{1(✉)}, Pyke Tin¹, and Hiromitsu Hama²

¹ University of Miyazaki, Miyazaki, Japan

thithi@cc.miyazaki-u.ac.jp, pyketin11@gmail.com

² Osaka City University, Osaka, Japan

hama@ado.osaka-cu.ac.jp

Abstract. In today world a new buzzword Internet of Things has been on the news nearly every day. Some researchers are even using Internet of Every Things. Its potentialities and applicability are now on the cutting edge technology. Also, all most all of business, health care, academic institutions are in one way or another, having to deal with the Internet of Things. So the Internet of Things reliability becomes an important factor. In this paper we proposed a Markov Queuing approach to analyze the Internet of Thing reliability. Since queuing theory investigates the delay and availability of functioning things and Markov concepts take the dependency of Things in the Internet, the combination of these two concepts will make the problem clear and soluble. For illustration, we present some experimental results.

Keywords: Internet of Things · Queuing theory · Markov dependency · Reliability · Availability

1 Introduction

Now, as an era of Internet of Things (IoT) all most all of things, human beings, animals, objects (things) are being and will be connected over the internet. Therefore, IoT can be thought of an advanced version of Machine to Machine communication system in which an object here you can say thing connects to another without human interventions. For example, when you go to a hospital, once you pass the registration counter, the machine scan your registration card and make instructions to the unit where you may have medical treatment and the registration. One thing we would like to point out is that everything in IoT has the expected life time which leads to the question of how much the things are reliable during their life time in IoT [1].

According to recent surveys, it is learnt that the number of connected IoT devices estimated 212 billion by the year 2020 creating the amount of data to reach the level of 40 ZB [2, 3]. Thus, the advancement of new technologies to analyze and storing such huge big data is a challenging task on demand. The number of applications based on IoT is fast growing and will be increasing more over the next few years. Generally speaking the functional components of Internet of Things include connected devices which send and receive data, application systems which are in the data centers or data followed by information queue linked between devices and application systems in IoT.

The information queue has been widely used to control the huge amount of data created by the IoT. As the name suggests that the main function of information queue is to make scheduling messages for the IoT working smoothly [4].

The general architecture of the Internet of Things consists of an IoT access network along with Internet protocol (IP), information queue systems which are linked to IoT application networks. Usually, the IoT devices generate the information in the access networks and transmit to information queuing systems where they are to be analyzed and processed by using some techniques of data analysis. Then, the queue services make decisions based on the processed information to give commands to actuators in the application network [5–9]. Thus the IoT services include functions of monitoring data in the physical domain of the access network, decision making functions in the information queues for application networks. Here, it is worthwhile to note that the concepts of queuing theory play important roles in dealing with the process of IoT devices and related information decision makings. Since queuing theory is involved there are the arrival process of information or messages and departure process of messaged will be necessary to be thoroughly investigated to make correct decisions. For this concerns, the way how the information are arriving that is what will be a suitable probability distribution for the aril process to be used and what kind of execution services will be given to the arriving information to the queues are important factors for making IoT services reliable and efficient. However we should realize that some IoT devices may be failure to produce correct and accurate information. In such cases, the information will produce wrong decisions so that the whole system can be put at the state of unreliable and inefficient service. Therefore the reliability measures of IoT devices needed to be automatically detected by the system. Otherwise, those unreliable devices will continue producing ill and inaccurate data so that the whole information queue will be making wrong decisions continuously. As an old concept at an early stage of using computers, we will have the results of “garbage ins garbage outs”. But these kinds of failures are curable by making system for measuring reliability of devices and information they generated to have the whole system be granted by accurate and reliable data [10]. To measure the reliability of IoT devices we will employ the theory of queue system which can investigate the message delays, traffic congestion and busy times. By using those quantities, we can find the reliability measures for the Internet of Things.

In this paper we shall propose the queuing theory approach by adding the Markov Chain concept to investigate some problems of IoT reliability. In the following we organize the rest of the paper as some related works in Sect. 2, the overview of proposed system in Sect. 3 and some illustrative experimental results in Sect. 4 followed by conclusions in the final section.

2 Some Related Works

In this section, we shall present some related works to this paper. Generally, most of IoT devices are based on the wireless communication system in which the data trans-missions are to be made among each other. We notice that channel errors are higher in wireless communication than those in wired communication. As a

consequence, the higher channel error can make the data damaged so that it will become data congestion by a heavy traffic. Such damaged data cannot be well organized for information processing and analysis so that the integrity of data cannot assured for further processing. Once the integrity of data is not ensured, the products of the system are not reliable. So we should exclude the damaged data and unreliable data from the system processing. In order to detect the damaged data or unusual data, an intrusion detection system can be employed [11, 12]. Generally, usual or normal data in IoT is periodically generated. Sometimes, data are derived based on events and such data are known as event driven data which do not belong to the category of normal data concept. However we should not ignore the event driven data for inputting into the information queue because of they are not normal data. So much for data concepts. Let us consider the problem of IoT service performance modeling techniques. For this purposes, we can model the quality of service measures as a function of response time, throughput and network utilization. Some aspects of such models have been appeared in the literature [13–15].

Among them some authors proposed and validated web application performance tool for the performance based on the response time distribution of IoT cloud system modeled on a classic M/M/m queue open network by assuming an exponential density function for the inter-arrival and service times. Using the response time distribution, they determined the optimum level of service and the relationship between the maximum number of tasks and the minimum number of resources (virtual machines). The response time takes into account both the waiting time in the queue and the service time. However, the existing ways of data processing is not enough for providing reliable IoT services. Therefore, in this paper, we propose a Markov dependent queuing system for information queue which can adaptively finds the suspicious data using the waiting times and expected queue length. However, while we are dealing with the arrival process for the information queue, we employ the distributions of two independent variables representing the normal data and event driven data only. The damaged data are not reliable to be used so we omit the damaged data in the inflows processing.

3 Proposed Architecture for Internet of Things Reliability

In this section, we propose a Markov dependent queuing model based the internet of things reliability measures by using response times, waiting times and busy times the IoT data transmitted from the various types of IoT devices. The proposed model architecture can be seen as described in Fig. 1. Specifically the IoT network has a multiple entry points which are something like multiple servers in a queue. Let the number of entry points or servers be N for $i = 1, 2, \dots, N$. Thus the message queue in the IoT can be seen as a multiple server queuing system where the information transmitted by the IoT devices which are processed for the next step to the application layers.

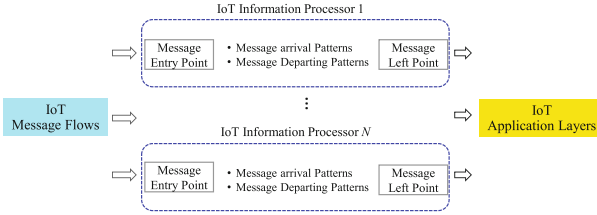


Fig. 1. Architecture of IoT message queue

To be specific, we assume that the transmitted information from the IoT devices join the message queue according to Markov distribution described in Eq. (1).

$$g(x) = \begin{cases} \frac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)}, & x > 0 \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

with $E(x) = \frac{\alpha}{\lambda}$, $Var(x) = \frac{\alpha}{\lambda^2}$.

Each processing unit in message queue has the same service rate and is equal to μ , this is $\mu = \mu_i$, $i = 1, 2, \dots, N$. Connecting servers in the message queue are fed with gamma arrival and service distributions. So α/λ is the feeding distribution for the customers leaving the entering server.

Fundamentally, a queuing system is a special type of Birth-Death process in Markov Chain theory in which the arrival in a queue means birth and the service means death. In this terminology, let the number of population (here in IoT the number of packets of messages) in the system (message queue) be n , then we can define λ_n and μ_n as the arrival and service rates depend on the number in the system.

Let $Q(t)$ be the number customers in the system at time t . Define

$$P_n(t) = \Pr[Q(t) = n | Q(0) = i]. \quad (2)$$

This will lead to the general forward Kolmogorov equation for $p_n(t)$ as described in Eq. (3).

$$\begin{aligned} \frac{dP_0(t)}{dt} &= -\lambda_0 P_0(t) + \mu_1 P_1(t), \\ \frac{dP_n(t)}{dt} &= -(\lambda_n + \mu_n) P_n(t) + \lambda_{n-1} P_{n-1}(t) + \mu_{n+1} P_{n+1}(t), \quad n = 1, 2, \dots \end{aligned} \quad (3)$$

In a state of equilibrium, also known as steady state, the behavior of the process is independent of the time parameter and the initial value; i.e.

$$\lim_{t \rightarrow \infty} P_n(t) = p_n \text{ when } t \rightarrow \infty, \quad n = 0, 1, 2, \dots$$

and therefore $\frac{dP_n(t)}{dt} = 0$ as $t \rightarrow \infty$, using these results in equation (3), we get

$$\begin{aligned} 0 &= -\lambda_0 P_0 + \mu_1 P_1 \\ 0 &= -(\lambda_n + \mu_n) P_n + \lambda_{n-1} P_{n-1} + \mu_{n+1} P_{n+1}, \quad n = 1, 2, \dots \end{aligned} \tag{4}$$

Solving the Eq. (4) recursively, we get

$$P_n = \frac{\lambda_0 \lambda_1 \dots \lambda_{n-1}}{\mu_1 \mu_2 \dots \mu_n} P_0 \tag{5}$$

By using $\sum P_n = 1$,

$$p_0 = \frac{1}{1 + \sum_{n=1}^{\infty} \frac{\lambda_0 \lambda_1 \dots \lambda_{n-1}}{\mu_1 \mu_2 \dots \mu_n}} \tag{6}$$

To make the problem simple, for N servers queue we have $\lambda_n = \lambda, n = 0, 1, 2, \dots$

$$\mu_n = \begin{cases} n\mu, & 0 < n < N \\ N\mu, & n > N \end{cases}$$

Then, Eqs. (5) and (6) becomes,

$$\begin{aligned} P_n &= \begin{cases} \frac{1}{n!} \left(\frac{\lambda}{\mu}\right)^n P_0, & n \leq N \\ \frac{1}{N! N^{n-N}} \left(\frac{\lambda}{\mu}\right)^n P_0, & n > N \end{cases} \\ P_0 &= \frac{1}{1 + \sum_{n=1}^N \frac{1}{n!} \left(\frac{\lambda}{\mu}\right)^n + \sum_{N+1}^{\infty} \frac{1}{N! N^{n-N}} \left(\frac{\lambda}{\mu}\right)^n} \\ &= \frac{1}{\sum_{n=0}^{N-1} \frac{1}{n!} \left(\frac{\lambda}{\mu}\right)^n + \frac{1}{N!} \left(\frac{\lambda}{\mu}\right)^N \left(\frac{N\mu}{N\mu - \lambda}\right)} \end{aligned}$$

Hence the probability of a new arrival waiting is given by

$$\begin{aligned} \text{Pr}(waiting) &= \sum_{n=N}^{\infty} P_n = \sum_{n=N}^{\infty} \frac{1}{N! N^{n-N}} \left(\frac{\lambda}{\mu}\right)^n P_0 = \frac{\frac{1}{N!} \left(\frac{\lambda}{\mu}\right)^N \left(\frac{N\mu}{N\mu - \lambda}\right)}{\sum_{n=0}^{N-1} \frac{1}{n!} \left(\frac{\lambda}{\mu}\right)^n + \frac{1}{N!} \left(\frac{\lambda}{\mu}\right)^N \left(\frac{N\mu}{N\mu - \lambda}\right)} \tag{7} \\ &= c(N, \mu, \lambda) \end{aligned}$$

The expression in (7) is also known as Erlang's C formula or Erlang's delay formula. By using this formula, we can compute the waiting time distribution as follows.

$$\Pr(W > t) = c(N, \mu, \lambda)e^{-\left(\frac{N\mu}{N\mu - \lambda}\right)} \quad (8)$$

This probability can be considered as IoT device failure if the t is big enough. Therefore, we define the reliability function $R(t) = 1 - \Pr(W > t)$. By using Eq. (8), we finally obtain the reliability function for each server in IoT message queue becomes as described in Eq. (9).

$$R(t) = 1 - c(N, \mu, \lambda)e^{-\mu\left(\frac{N-\lambda}{\mu}\right)t} \quad (9)$$

That can be interpreted as the probability of a sever functioning until time t . Thus the probability of the whole IoT message queue to have reliable informing processing can be determined by using a threshold value say α such that $P = 1 - [1 - R(t)]^N \geq \alpha$, otherwise the IoT system is not reliable.

4 Some Simulated Experimental Results

In order to make an illustration for the validity of the proposed method we compute the reliability measures for the IoT devices and the information generated. In doing so, the correlation between the IoT devices and the number that describe the availability of quality services are derived by using numerical methods. In current status, the most of existing methods for measuring the reliability of information utilized some kinds of iterations processes. Generally speaking, the methods of iterations were mainly involved with the concepts of convergence. However, the proposed method in this paper does not require the convergent criteria so that it makes easier for application with respect to computing. Thus in calculating the reliability measures of IoT services and devices, the computing times are significantly reduced. The fluctuations of reliability distribution patterns are calculated by using queuing theory concept proposed in Sect. 3. We present some numerical results to show the confident levels of parameters such as the traffic intensity and the correlation between the devices.

In Table 1, we present some results of reliability measures for IoT devices for large correlation values and small size of demands. On the other, Table 2 shows the reliability results by using minimum correlation values and maximum input size. The corresponding results are also expressed in the form of graph in Figs. 2 and 3. Again, through the varying the parameters thresholds, inputs and the number of devices it is observed that the threshold value is range from 0.7 to 0.9 for large value of N , the total number of IoT things within the datasets.

Table 1. Distribution of reliability patterns for small $N = 10$

N	1	2	3	4	5	6	7	8	9	10
S_1	0.33	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
S_2	0.40	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
S_3	0.43	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
S_4	0.44	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
S_5	0.45	0.99	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 2. Reliability counts for variable thresholds

N/α	1	2	3	4	5	6	7	8	9	10
0.7	0	5	5	5	5	5	5	5	5	5
0.75	0	5	5	5	5	5	5	5	5	5
0.8	0	5	5	5	5	5	5	5	5	5
0.9	0	5	5	5	5	5	5	5	5	5

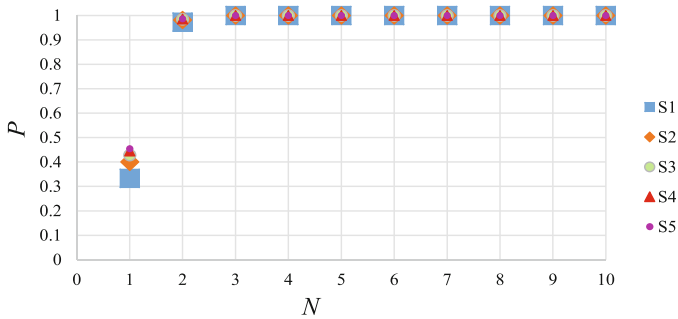


Fig. 2. Analysis of P reliability probability on each parameter set

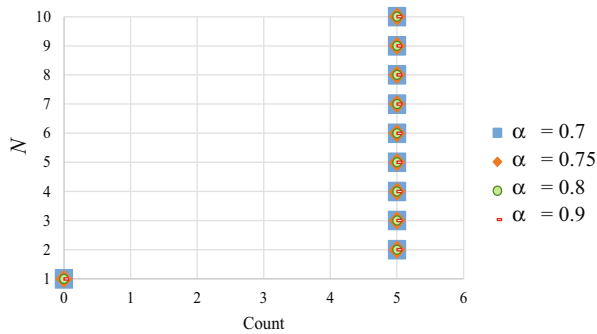


Fig. 3. Variation of thresholds for reliability

5 Conclusions

In this paper we had presented a new look into the reliability measuring system for the Internet of Things environments. We observed that the approach based on the Markov queuing theory seems to be efficient and promising. Although we have employed the synthetic data for illustration, we will use real world datasets in future.

Acknowledgments. This work was supported in part by SCOPE: Strategic Information and Communications R&D Promotion Program (Grant No. 172310006) and JSPS KAKENHI Grant Number 17K08066.

References

1. Lu, R., Li, X., Liang, X., Shen, X., Lin, X.: GRS: The green, reliability, and security of emerging machine to machine communications. *IEEE Commun. Mag.* **49**(4) (2011)
2. Gantz, J., Reinsel, D.: The digital universe in 2020: big data, bigger digital shadows, and biggest growth in the far east. IDC iView: IDC Analyze Future **2007**(2012), 1–16 (2012)
3. Diaz, M., Martín, C., Rubio, B.: State-of-the-art, challenges, and open issues in the integration of Internet of Things and cloud computing. *J. Netw. Comput. Appl.* **67**, 99–117 (2016)
4. Buyya, R., Dastjerdi, A.V. (eds): *Internet of Things: Principles and Paradigms*. Elsevier, Amsterdam (2016)
5. Kim, S., Na, W.: Safe data transmission architecture based on cloud for Internet of Things. *Wireless Pers. Commun.* **86**(1), 287–300 (2016)
6. Kang, Y.: New approach to the platform for the application development on the Internet of Things environment. *J. Platform Technol.* **3**(1), 21–27 (2015)
7. Gubbi, J., Buyya, R., Marusic, S., Palaniswami, M.: Internet of Things (IoT): a vision, architectural elements, and future directions. *Future Gener. Comput. Syst.* **9**(7), 1645–1660 (2013)
8. Mahamure, S., Railkar, P.N., Mahalle, P.N.: Communication protocol and queuing theory-based modelling for the internet of things. *J ICT Stand.* **3**(2), 157–176 (2015)
9. Kim, D.Y., Jeong, Y.S., Kim, S.: Data-filtering system to avoid total data distortion in IoT networking. *Symmetry* **9**(1), 16 (2017)
10. Chelli, K.: Security issues in wireless sensor networks: attacks and countermeasures. *Proc. World Congr. Eng.* **1**, 1–3 (2015)
11. Sen, J.: A survey on wireless sensor network security. *Int. J. Commun. Netw. Inf. Secur.* **1**, 55–78 (2009)
12. Albers, P., et al.: Security in Ad Hoc networks: a general intrusion detection architecture enhancing trust based approaches. In: *Wireless Information Systems*, pp. 1–12 (2002)
13. Guo, L., Yan, T., Zhao, S., Jiang, C.: Dynamic performance optimization for cloud computing using M/M/m queueing system. *J. Appl. Math.* (2014)
14. Railkar, P.N., Mahalle, P.N.: A queuing theory-based modelling for performance analysis towards future internet. In: *Annual IEEE India Conference (INDICON)*, pp. 1–7 (2014)
15. Vilaplana, J., et al.: A queuing theory model for cloud computing. *J. Supercomput.* **69**(1), 492–507 (2014)

Some Characteristics of Nanyaseik Area Corundum and Other Assorted Gemstones in Myanmar

Htin Lynn Aung¹ and Thi Thi Zin²(✉)

¹ Department of Geology, Bago University, Bago, Myanmar
htinlynnang119@gmail.com

² University of Miyazaki, Miyazaki, Japan
thithi@cc.miyazaki-u.ac.jp

Abstract. The rock sequence of the study area consists of medium to high grade metamorphic rocks, marble, gneiss and intrusive igneous rocks, mainly biotite microgranite and serpentinite. Although the primary occurrences of gemstones in this area seem to be scarce, the secondary placers gem-bearing deposits are noteworthy. The Nanya rubies are characterized by their distinct colours of which, the commonest colour being, light pinkish red and intense red and rarely pigeon's blood red. A glassy texture with excellent transparency makes the stone more attractive. In crystal forms, rubies usually have rounded corners, rhombohedrons, pinacoids and not well developed prism faces. Habitually, rhombohedral faces display coarse striations and some with pitted surfaces. It is probable that the Nanyaseik area is situated near the plate boundaries and within the northern splay of the Sagaing fault. Moreover, it also forms a segment of Jade Mine region. Therefore, it is reasonable that the pressure had played an important role more effectively than the temperature in the process of metamorphism.

Keywords: Nanyaseik area · Secondary deposits · Hpakant Township · Byones · Jade Mine region

1 Introduction

Myanmar is rich in mineral products (Chhibber 1934a, b, Clegg (1941)). Among them, gemstones have some special characteristics in color as well as in textures. In this study, we present some finding of practical field works in the area of Nanyaseik located in Kachin State of Myanmar (DGSE 1995). Specially, we investigate the identification of corundum and other assorted gemstones and the quality of gemstones in comparison with those of Mogok Stone Tract (Nu 2003).

2 Methods of Study

Minerals and gems sampling for detailed mineralogical and gemological are studied (GIAC (1999), Aung (2000)). Identification of corundum and other precious gemstones were carried out with the aid of gemmolite, microscope, refractometer, polariscope,

spectroscope and other available gem testing instruments as shown in Table 1 (Figs. 1, 2, 3, 4, 5, 6, 7, 8 and 9; Tables 2, 3, 4, 5, 6, 7, 8 and 9).

Table 1. Gemmological studies of corundum and other assorted gemstones

Corundum	
Chemical composition	Al ₂ O ₃ Aluminium oxide
Crystal system and habits	Trigonal. Prismatic or tabular six-sided crystals, often with flat basal terminations. Rhombohedron faces may be developed at alternate corners. Also occurs as six-sided bipyramids, with varying angles
Cleavage	A false cleavage, or ‘parting’ of twin planes occur parallel to the basal and rhombohedral faces
Hardness	9
SG	3.9–4.1
Colours and varieties	Red—ruby, blue—sapphire, orange-pink—padparadscha, other colours are called coloured sapphires, e.g. green sapphire and yellow sapphire
Lustre	Vitreous to bright vitreous
RI	1.760–1.768
Birefringence	0.008–0.009
Optical nature	Uniaxial –ve
Dispersion	Low
Optical effects	Asterism, six-rayed stars (rarely twelve-rayed)
Inclusions	Fine rutile needles, termed ‘silk’ when numerous; ‘feathers’—partly healed fractures; twin planes, zircon haloes, growth zoning and other solid (minerals)

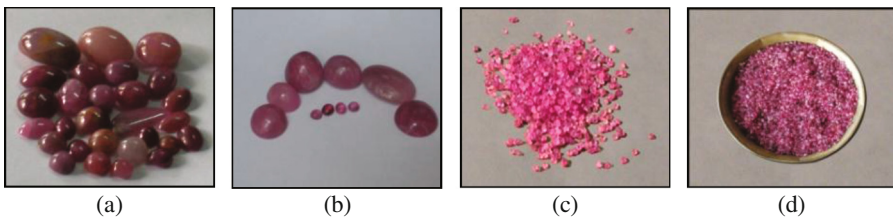


Fig. 1. (a, b) Cabochon cut rubies, (c, d) Natural uncut rubies



Fig. 2. (a) Sapphire crystal (rough), (b) Rough crystals of sapphire, (c) Cabochon cut sapphires



Fig. 3. (a) Padparadscha (rough), (b) Padparadscha (mixed cut)

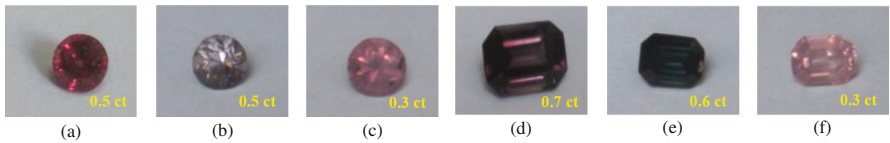


Fig. 4. (a, b, c) Mixed cut spinels (various colours), (d, e, f) Emerald cut spinels (various colours)

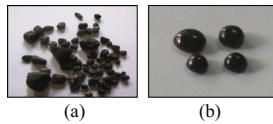


Fig. 5. (a) Tourmaline crystals (schorl), (b) Tourmaline cabochons (schorl)

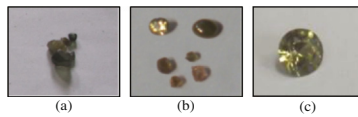


Fig. 6. (a) Zircon crystals, (b) Zircon (brilliant cut, cabochon and rough crystals), (c) Zircon (brilliant cut)



Fig. 7. (a) (i) Smoky quartz, (ii) and (iii) Rock crystals (nearly water worn), (b) Yellowish colour quartz

Almandine garnet	
Chemical composition	$Fe_3Al_2(SiO_4)_3$ Iron aluminium silicate.
Colour	Purplish-red, pale to deep mauve.
Hardness	7½
SG	3.8 to 4.2
Absorption Spectrum	Absorption is due to iron. Three broad bands in the yellow, green and blue-green are prominent.
Lustre	Bright vitreous.
RI	1.760 – 1.790
Inclusions	Rounded or irregular 0.2 ct crystal inclusions.
Fashioning	Cabochon.

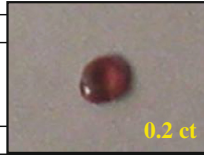


Fig. 8. Almandine garnet (cabochon cut)

DIOPSIDE	
Chemical composition	$CaMg(SiO_3)_2$ Calcium magnesium silicate.
Crystal system and habit	Monoclinic, prismatic crystals. Also occurs as water-worn pebbles.
Cleavage and fracture	Distinct prismatic; in two directions at almost 90°.
Hardness	5½
SG	3.26 to 3.32
Colour and varieties	Green to brown. Chrome diopside – a bright green variety in which dark green colours is partly due to chromium.
Pleochroism	Weak to moderate, two shades of the body colour.
Optical nature	Biaxial +ve.
Lustre	Vitreous.
Dispersion	Low.
Some distinguishing features:	
RI	1.675 - 1.701
Birefringence	0.2 ct 0.026
Fashioning	Cabochon.



Fig. 9. Diopside (cabochon)

Table 2. Gemmological studies of ruby

Ruby	
Colour	Light pinkish red (common), intense red or nearly Pigeon's blood red (rare)
Pleochroism	Strong to slight pink-red and orangish-red, dichroic
Localities	Ruby is found in commercial quantities in many locations of Nanyaseik, the most important of which are: Manaw, Sabaw, Lakha, Warphu and Khaing Kyin Worksites
Occurrence	Mostly in metamorphic and alluvial gem gravel deposits. Commercial production is mainly from alluvial gem gravel (placer) deposits
Fashioning	Cabochon

Table 3. Gemmological studies of sapphire

Sapphire	
Colour	Blue, pink, orange and green
Pleochroism	Strong in varieties other than yellow. Other colours show differing shades of the body colours. Blue sapphires show blue plus green, dichroic colours
Localities	Nanyaseik
Occurrence	Production mostly from placer gravels
Fashioning	Cabochon

Table 4. Gemmological studies of Padparadscha

Padparadscha	
Hardness	9
SG	3.97–4.01
Colour	Orange-pink (padparadscha)
Lustre	Vitreous to bright vitreous
RI	1.761–1.770
Birefringence	0.009
Optical nature	Uniaxial –ve
Dispersion	Low
Inclusions	Fine rutile needles, termed 'silk' when numerous; 'feathers'—partly healed fractures; twin planes, zircon haloes, colour and growth zoning
Fashioning	Mixed cut

Table 5. Gemmological studies of spinel

Spinel	
Chemical composition	MgAl ₂ O ₄ magnesium aluminium oxide
Crystal system	Cubic
Cleavage	None. Most spinels are brittle
Hardness	8
Specific gravity	3.58–3.61
Colours and varieties	Spinel is allochromatic and exhibits a very wide range of colours, red, pink, light purple, dark green, orangy pink. Most apparently colourless stones are a very pale shade of pink or lilac. Chromium imparts a pink or red colour
Lustre	Vitreous, bright. Some crystals have smooth surfaces with bright vitreous luster
Refractive index	1.715–1.720. Single
Inclusions	Many natural spinels contain minute octahedral crystals which may be other spinel-type minerals. Zircon haloes are also seen. Iron-stained fractures are common
Localities and occurrence	Most gem spinels occur in alluvial deposits, together with corundum (in Nanyaseik)
Fashioning	Mixed cut, brilliant cut, emerald cut (step cut) and cabochon

Table 6. Gemmological studies of tourmaline

Tourmaline	
Chemical composition	Complex boro-silicate of aluminium, magnesium, iron, calcium and alkali elements
Crystal system and habit	Trigonal, three sided prisms; the faces of triangular prisms are often convex, resulting in a rounded triangular cross section with vertical striations
Cleavage and fracture	Fractures frequently seen as wavy cracks in rough stones near perpendicular to the c-axis
Hardness	7–7½
SG	3.06
Colour	Black and dark green
Lustre	Vitreous
Dispersion	Low
Optical effects	Chatoyancy (caused by parallel fibres or tubes)
<i>Some distinguishing features</i>	
Pleochroism	Strong in most stones, depending on colour and depth of colour. Some green stones show brown/green or even black/green. However, a few show no pleochroism
RI	1.635–1.672
DR	0.025
Optical nature	Uniaxial -ve
Inclusions	Irregular or wavy partially healed fractures and fluid inclusions.
Fashioning	Cabochon

Table 7. Gemmological studies of Zircon

Zircon	
Chemical composition	ZrSiO ₄ Zirconium silicate. Zircon may also contain minor amounts of other elements including uranium and thorium
Crystal system and habit	Tetragonal. Crystals are prismatic, square in cross section and terminated by pyramids. Zircon also occurs as water worn pebbles
Hardness	High type: 7½. Zircon is brittle, so that facet edges are easily damaged; chipping can occur if loose stones are allowed to rub together in a stone paper. For this reason, cut zircons are often individually wrapped in tissue paper
SG	4.8 (high)
Colour	Yellowish green and brown
Pleochroism	Weak, except in heat-treated blue stones whose dichroic colours are blue and colourless
RI	High type: 1.92–1.99
Optical nature	Optical Nature: uniaxial +ve, but metamict stones may be almost isotropic
Lustre	Sub-adamantine to bright vitreous
<i>Some distinguishing features</i>	
Absorption spectrum	653.5 nm and 659.0 nm in the red is diagnostic for zircon
Birefringence	A maximum of 0.059 (high zircon) to almost none in metamict stones
Dispersion	High
Observation	Doubling of the back facets, which is often easily visible with a 10x lens. Bright luster; damaged facet edges
Fashioning	Brilliant cut, cabochon

Table 8. Gemmological studies of quartz

Quartz	
Chemical composition	SiO ₂ Silicon dioxide (also known as silica)
Crystal system and habit	Trigonal. Hexagonal prismatic crystals with horizontal striations and rhombohedral sets of pyramidal terminations which usually look like hexagonal pyramids
Cleavage	Very poor. Fracture generally conchoidal
Hardness	Crystalline: 7
SG	Crystalline: 2.65
Lustre	Vitreous
RI	Crystalline 1.544–1.553
Birefringence	0.009
Optical nature	Uniaxial +ve

Table 9. A comparison of gem occurrences of Nanyaseik and Mogok Byones (modified after Win et al. (2004))

	Mogok gem occurrence	Nanyaseik gem occurrence
Bed rock	Marbles and granitoid rocks	Basement, gneiss rocks overlain by alluvial sediments
Physiography	Rugged mountainous terrain	Opened flat plain with few barriers of localized hills and ranges
Elevation	Over 3800' above sea-level	About 550' above sea-level
Type of basin	Almost closed type	Mostly open pit (Inn Bye) method
Gem mining area	Over 12 × 5 miles	About 5 × 5 miles
General basinal slope of soil formations	Fair to steep slope	Fair to gentle slope
Byone associated soil type	Commonly transported soils, rarely residual soils	Generally lateritic soils
Depth of byone	Ten feet to over a few hundred feet	About four feet to under forty feet
Depth of bed rock	Very variable a few to hundreds of feet	Mostly about eighty-five feet
Lateritization	Unusual	Usually common
Karst topographic features	Fair to well developed	Not common, except Khaing Kyin Taung
Associated gem varieties	Rubies, sapphires and many other precious stones	Mainly ruby, spinel, sapphire (padparadscha), and other precious stones
Gold association	Uncommon	Localized

3 Conclusion and Future Prospects

All gemstone occurrences from Nanyaseik area are mainly recovered from secondary deposits (gravels) (Theory and Practical Hand Book, Aung (2004), Webster (1970), Sann (2010)). These are transported, deposited and accumulated in the adjacent valleys and flat lowland areas. Ruby, sapphire, spinel, zircon, tourmaline, quartz, garnet, diopside and other precious stones are extracted from byones, gem bearing gravels (Hutchison (1975), Arem (2010), Hurrel, Htay (2010) Thin (1991)).

This research work is not the end of the story of gems occurrences of Nanyaseik Stone Tract. Some more concealed gemstones could be discovered in the near future by local people, geologists, gemologists, gems dealers, and gems collectors.

In this paper, we have discussed the characteristics of gemstones with respect to geological terms and terminologies. However, in future we shall investigate the geological findings by using image processing techniques. Since the gemstones are mainly dependent on their distinct color and texture, the image technology can make more promising results.

References

- Chhibber, H.L.: The Geology of Burma, p. 538. Macmillian & Co. Ltd., London (1934a)
- Chhibber, H.L.: The Mineral Resources of Burma, p. 320. Micmillian & Co. Ltd., London (1934b)
- Clegg, E.L.G.: The cretaceous and associated rocks of Burma. Mem. Geol. Surv. India. **74**, 101 (1941)
- DGSE Staff Report, Gemstone Occurrences of Nanyaseik Area, Kamaing Township, Kachin State. No. 4, p. 8 (1995)
- GIAC Final Report, GIAC Project, TOTAL-UNOCAL-MOGE, Ecole Normale Superieure, Yangon, Dagon, Mandalay, Chiang Mai and Chulalongkorn Universities (1999)
- Aung, H.H.: Mineralogy and Petrology of the Balon-Pingu Taung Area (with emphasis on genesis of ruby and sapphire). Unpubl. M.Sc. Thesis, University of Yangon (2000)
- Win, H., et al.: Ruby and Sapphire occurrences of Nanyaseik, Kachin State, Northern Myanmar. J. Asia Res. Centre **2**, 18 (2004)
- Hutchison, C.S.: Ophiolite in Southeast Asia. Geol Soc. Am. Bull. **86**, 797–806 (1975)
- Arem, J.E.: Color Encyclopedia of Gemstones (2010)
- Hurrel, K., Johnson, M.L.: Gemstones (A complete color reference for precious and semiprecious stones of the world)
- Htay, K.N.: Geology and Occurrences of Jadeite Jade in Phakant, Lonekhin and Tawmaw Areas. Myitkyina District, Kachin State, Unpubl. MRes Thesis, University of Yangon (2010)
- Thin, N.: Tectonic environment of Jadeite Deposits of the Phakant-tawmaw area. Kachin State. Upper Myanmar Georeport **1**(1), 49–60 (1991)
- Nu, T.T.: A Comparative Study of the Origin of Ruby and Sapphire in the Mogok, Pyinlon and Mong Hsu Areas. Ph.D. Thesis, Mandalay University, Unpub Paper (2003)
- Theory and Practical Hand Book. Fellow of the Gemological Association and Gem Testing Laboratory of Great Britain (FGA)
- Aung, T.: Geology of the Nanyaseik Area, Mogaung Township, Myikyina District. Unpubl. M. Sc. Thesis, University of Yangon (2004)
- Webster, R.: “Gems”, Their Sources, Descriptions and Identification, 2nd edn, p. 836. Butterworths & Co. Publishers Ltd., Great Britain (1970)
- Sann, Z.O.: Petrology and Genetic Aspects of Gem Minerals in the KyetsaungTaung-Kyaukkyi Area, Thabeikkyin Township. Unpubl. Ph.D Thesis, University of Mandalay (2010)

Exploring Gemstones in Northern Part of Myanmar

Htin Lynn Aung¹ and Thi Thi Zin²(✉)

¹ Department of Geology, Bago University, Bago, Myanmar
htinlynnaung119@gmail.com

² University of Miyazaki, Miyazaki, Japan
thithi@cc.miyazaki-u.ac.jp

Abstract. The primary occurrences of gemstones in Nanyaseik area seem to be scarce, the secondary placer gem-bearing deposits are noteworthy. All gemstone occurrences from Nanyaseik area are mainly recovered from secondary deposits (gravels). Gemstones are found as detrital fragments in gem-bearing soil horizons known as byones. According to the drainage characteristics of this area and its environs gem-bearing alluvium had been probably descended from north-western and western watersheds that created those secondary deposits, especially at the junctions of major streams and their tributaries where local people wash the byone and extract gems. These gems include precious rubies, sapphires (including padparadscha) and others; spinel, tourmaline, zircon, quartz, diopside and almandine garnet.

Keywords: Nanyaseik area · Secondary deposits · Phakant township · Byones · Kachin state

1 Introduction

In Myanmar Universities, the discipline of geological studies involved field works in practical areas of mines. In this aspect, gemology is one of important studies because Myanmar has been famous due to the products of precious gemstones. In this paper, we investigate the occurrences of gemstones in Phakant area in the northern part of Myanmar. This region is covered about 130 km² of mountains and rugged terrain (see Fig. 1). Generally, the gem materials form minerals in which ruby, sapphire, spinel, zircon, tourmaline, quartz, diopside and almandine garnet can be found.

In this study, we shall investigate three aspects of gemology in Kachin state namely: (i) study the gem occurrences in Nanyaseik area, (ii) carry out laboratory works including mineralogical studies and identification of gemstones, and (iii) study the quality of gemstones in comparison with those of Mogok Stone Tract in the hope of gaining attention of mineralogists, gemologists, geologists, other gem dealers and gem collectors.

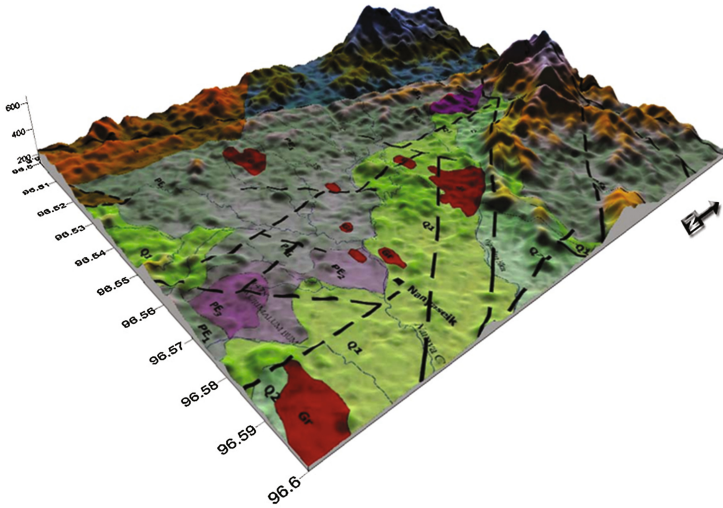


Fig. 1. 3D View of geomorphology of the Nanyaseik area

2 Mining Methods

Mechanisation is used wherever possible to ensure profitable yield. Mining methods applied in Nanyaseik area are as follows: (i) Open pit mining method (Inn Bye) [Fig. 2 (a)], (ii) Square pit mining method (Lebin Twin) [Fig. 2(b)].



Fig. 2. Mining methods (a) Open pit mining method (Inn Bye), (b) Square pit mining method (Lebin Twin)

2.1 Open Pit Mining Method (Inn Bye)

This method is mainly applied on secondary deposits. Almost all the mining activities in Nanyaseik area are carried out by means of this method for voluminous deposits. The mining operations involve: (a) Removing the overburden, (b) Excavation of gem bearing gravel, and (c) Sluicing techniques:

- (a) *Removing the overburden*: the removal of overburden is carried out aided by bulldozers where the top soil is removed until gem bearing gravels (byones) exposed.
- (b) *Excavation of gem bearing gravel*: It is carried out by means of excavators.
- (c) *Sluicing techniques*: This sluicing method is very popular and carried throughout the year. It commonly applied on hillsides and steep valleys. It is an open trench method, local people called Myaw-dwin which are being sluiced by making use of water power that drains from the higher levels. The water supply is very important, conventionally by passing it along suitably cut bamboos or by water pipes operated by diesel engines. The top soil is first removed until the byone layer is exposed. The byones are then washed on a nearby flat circular floor. Water is diverted into it along with the byone in the form of sprays and the washed up materials enter the trailing canal and the heavy materials are trapped in pits or sluices.

2.2 Square Pit Mining Method (Lebin Twin)

There are about three thousands of old and current square pits at Nanyaseik area. It is a small square pit about four square feet. The depth may vary from (5–20) feet depending on the byone layer and bed rock.

Gem Worksites: In Nanyaseik area, the workable gem worksites are mostly situated in the western environs of Nanyaseik village. These worksites are generally N-S trending. Mine area extends about 8 km in NS, 5 km in EW directions. Based on the field evidences, rubies and other gemstones in Nanyaseik area are originated in marbles, and associated igneous rocks. However, gemstones are extracted from byones, secondary deposits (gravels) rather than from insitu (primary deposits). Notably sample gem worksites are outstanding and are listed below (Fig. 3). Essential products of these gem worksites are primarily ruby and other semi-precious stones like tourmaline, spinel, garnet, zircon, quartz, etc. Annually workable period is only in dry season because of the heavy rainfall and swampy conditions in rainy season. Some gemologists stated that rubies and other precious stones are obtained from the alluvial deposits north of the village which are more productive than those to the south. These gemstones were derived from the detritus, afforded by the disintegration of crystalline limestones adjacent to the intrusive masses of granite.

(a) *Sample Worksite 1*

It is situated in the northern most part of the study area. It uses a square pit mining method (Fig. 4). It is quite small, about three square feet, locally called Lebin twin. The depth reaches to twenty-five feet, producing only ruby and other gemstones. The workable period is lasts only about five months producing about thirty carats of gem quality stones.

(b) *Sample Worksite 2*

In this worksite, local people used an open pit mining technique (Inn bye method) (Fig. 5(a)). Basement rock (host rock) is NE dipping gneiss which shows reddish brown colour on weathered surfaces. The reddish brown soil cover is ten feet thick. Feldspars are mostly weathered and altered to clay (sticky mud). At this

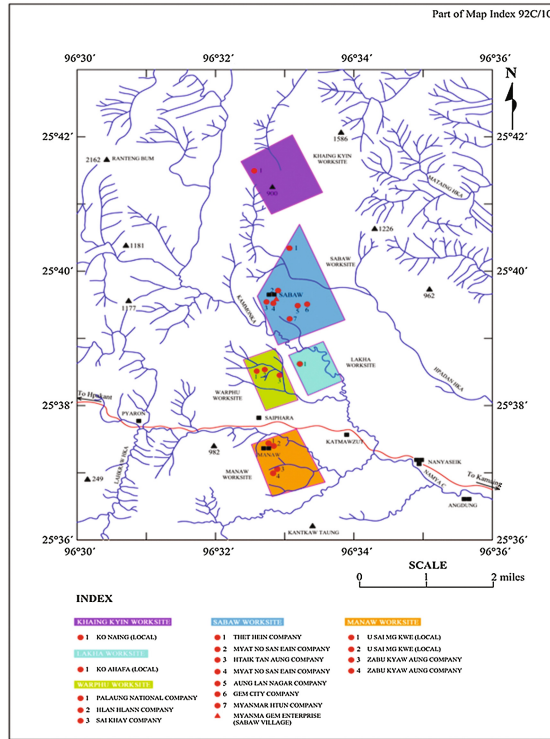


Fig. 3. Map showing gemstone workings in Nanyaseik area



Fig. 4. Sample Worksite 1: (a) A square pit (Lebin Twin), (b) Rubies sorted out from the byones

worksite, rubies and other associated gemstones are recovered from secondary placer deposits (Fig. 5(b)).

(c) *Sample Worksite 3*

This worksite lies in the central part of the Nanyaseik area.

(d) *Sample Worksite 4*

This worksite is situated in the southernmost part of the study area. Local people said that it was a famous worksite prior to the Worksite 3.



Fig. 5. Sample Worksite 2: (a) Recovery of byones by pulling process, (b) Sorted out gem quality rubies

(e) *Sample Worksite 5*

This Worksite is next to the Worksite 2. The working mining companies use open pit technique to extract rubies and other gemstones. Gold particles can be seen on wooden pan (Fig. 6).



Fig. 6. Sample Worksite 5: (a) Recovery of byones by water sluicing method, (b) Gold particles

3 Nature of Secondary Deposits (Gravels) and Gemstone Association

All gemstone occurrences from Nanyaseik area are mainly recovered from secondary deposits (gravels). Gemstones are found as detrital fragments in gem bearing soil horizons known as byones. The maximum depth of byone is very shallow, never exceeds one hundred feet. In some places, byones have been found within the reach of one-foot depth from the surface. The soil type of byone is dominantly residual soil (laterite) overlying marbles. The thickness of soil ranges from 50' to 70'. In Nanyaseik area, basic soil profile succession can be listed as follows: (i) Recent yellow to brown sandy and clayey soil mixed with gravels, (ii) Bluish soil with organic matters, (iii) Laterite and lateritic soil. Weathered laterite and lateritic soil are common to all gem worksites and are mostly formed as the lower most part with well-developed byone horizons mixed with fragments of parent rocks occasionally and other soil types. The thickness of residual soil varies from a few inches to (10–20) feet.

Most of the workable gem worksites in Nanyaseik area applied open pit mining method due to the swampy lowland with very gentle slopes (5°–10°). Moreover,

hydraulic sluicing, gravel pumping and vibration jiggling methods are also practiced to recover the gems which include ruby and other precious gemstones like sapphire, spinel, garnet, diopside, zircon, tourmaline, quartz, etc. from byones. Finally these gems are sorted out by hand with tweezers. Inferior quality gemstones are also encountered in byones, elsewhere in gem worksites within shallow depths. Economically important, thick byone formations have not yet been found. In some places, opaque rubies, carbolates, as the native miners call (due to its resemblance to carbolitic-soap) are also to be observed. According to DGSE's staff report (1995), they examined the Nanyaseik area and stated that rubies and some assorted gemstones were obtained from the detritus, disintegration of crystalline limestones surrounded by intrusive body. They described five major localities in the environs of Nanyaseik area and show the production rate of rubies with their respective worksites as shown in Table 1. Conclusively, topographic features of the study area have not been a favourable site to form considerable thickness of gem bearing gravels (byones) in Fig. 7.

Table 1. Production rate of rubies

No.	Name of worksite	Product rate (ct/cubic yard)	Reserve JV worksite	Permitted JV worksite
1	Sample Worksite 1	Test pit- 2.5, Loodwin-12.5	121 blocks	3 blocks
2	Sample Worksite 2	1.2	12 blocks	–
3	Sample Worksite 3	2.4	102 blocks	–
4	Sample Worksite 4	1.2	40 blocks	–
5	Sample Worksite 5	1.4	33 blocks	–
			308 blocks	

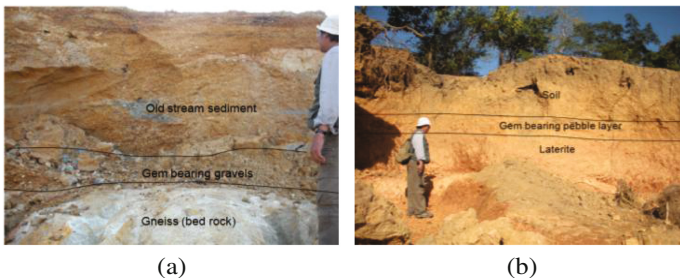


Fig. 7. (a) Gem bearing gravels sandwiched in between gneiss (bed rock) and old stream sediments, (b) Gem bearing pebbly layer sandwiched in between soil and laterite

4 Conclusion

All gemstone occurrences from Nanyaseik area are mainly recovered from secondary deposits (gravels). These are transported, deposited and accumulated in the adjacent valleys and flat lowland areas. Ruby, sapphire, spinel, zircon, tourmaline, quartz, garnet, diopside and other precious stones are extracted from byones, gem bearing gravels.

In future works, sorting and classification of gemstones will be done by using image processing techniques. We hope that the image based technology would be a promising approach because gemstones are mostly characterized by colors and textures appearances.

References

1. Bauer, M.: Precious Stones. Dover, New York (1968)
2. Bender, F.: Geology of Burma, p. 239. Gebruder Borntraeger, Berlin (1983)
3. Chhibber, H.L.: The Geology of Burma, p. 538. Micmillian & Co. Ltd., London (1934)
4. Chhibber, H.L.: The Mineral Resources of Burma, p. 320. Micmillian & Co. Ltd., London (1999)
5. DGSE Staff Report, Gemstone Occurrences of Nanyaseik Area, Kamaing Township, Kachin State, no. 4, p. 8 (1995)
6. Win, H., et al.: Ruby and Sapphire occurrences of Nanyaseik, Kachin State, Northern Myanmar. *J. Asia Res. Centre* **2**, 18 (2004)
7. Arem, J.E.: Color Encyclopedia of Gemstones. Van Nostrand Reinhold Company, New York (1987)
8. Thin, N.: Tectonic environment of Jadeite deposits of the Phakant-tawmaw area. Kachin State. *Upper Myanmar Georep.* **1**(1), 49–60 (1991)
9. Aung, T.: Geology of the Nanyaseik area, Mogaung Township, Myikyina District. Unpublished MSc thesis, University of Yangon (2004)

Encryption and Security

Attacks and Solutions of a Mutual Authentication with Anonymity for Roaming Service with Smart Cards in Wireless Communications

Tsu-Yang Wu^{1,2}, Bin Xiang³, Guangjie Wang³,
Chien-Ming Chen^{3(✉)}, and Eric Ke Wang³

¹ Fujian Provincial Key Laboratory of Big Data Mining and Applications,
Fuzhou 350118, China

wutsuyang@gmail.com

² National Demonstration Center for Experimental Electronic Information
and Electrical Technology Education, Fujian University of Technology,
Fuzhou 350118, China

³ Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, China
jiabinxiang@126.com, wgj13718925364@163.com,
Chienming.taiwan@gmail.com, wk_hit@hit.edu.cn

Abstract. Recently, Liu *et al.* proposed a mutual authentication protocol with user anonymity for wireless communication. In their paper, the authors claimed that the protocol can resist several kinds of attacks even the secret information stored in the smart card is disclosed. However, we still find two vulnerabilities in this paper. First, this protocol still fails to protect user anonymity. Second, this protocol is vulnerable to an off-line password guessing attack if an adversary can derive the secret information stored in a smart card. To solve the problems, we propose a simple but effective patch to their protocol.

Keywords: Anonymity · Roaming service · Wireless communication · Cryptanalysis

1 Introduction

Due to the rapid development of wireless networks, information exchanging between mobile devices (e.g., PDA, smart phone and laptop) have been enormously increased. Through the global roaming technology, if a mobile user roams into a foreign network, he needs to perform mutual authentication and establish a session with a foreign agent before transmitting secret information.

Recently, many authentication protocols [1–5] for the above environment have been proposed. In 2013, Guo *et al.* proposed a mutual authentication key agreement protocol [4] using a smart card for wireless communications. Their protocol is based on Chebyshev chaotic maps; thus, it is more efficient than previous works. However, in 2016, Liu *et al.* pointed out that Guo *et al.*'s protocol cannot resist an impersonation attack if an adversary derives the information stored in the smart card. In order to

defense such attacks, Liu et al. proposed a new protocol [5] based on quadratic residue. Liu *et al.* demonstrated that their protocol can provide user anonymity and is secure against several kinds of attacks even if the smart card is disclosed. Unfortunately, in this paper we find that Liu *et al.*'s protocol still fails to protect user anonymity. Besides, if the adversary can extract the secret information stored in a victim's smart card, he can carry out an off-line password guessing attack. In order to solve the drawbacks we found, we provide a simple but effective patch.

2 Review of Liu *et al.*'s Protocol

In this section, we review Lie *et al.*'s protocol [5]. This protocol contains three phases, the registration phase and the mutual authentication phase and the updated password phase. Notations used in this paper are listed in Table 1.

Table 1. Notations used in this paper

Notations	Description
HA	Home Agent of a mobile user
FA	Foreign Agent of the network
MU	Mobile User
PW_M	A password of MU
ID_X	Identity of an entity X
T_X	Time stamp by an entity X
K_{FH}	The pre-shared key between HA and FA
p_x, q_x	Large primes numbers
d	The secret key of HA
$h(\cdot)$	A one-way hash function
\parallel	String concatenation operation
\oplus	XOR operation

In the Liu *et al.*'s protocol, modular square root (*MSR*) [6–8] is utilized to ensure the security. More specifically, both Home Agent (HA) and Foreign Agent (FA) require to initialize some parameters before the authentication phase. HA selects two different large primes p_1, q_1 that satisfy $p_1 = q_1 = 3(mod 4)$ and computes $n_1 = p_1 * q_1$, it then chooses its secret key d . After that, it publishes n_1 and the one-way hash function $h(\cdot)$. FA also selects two distinct large primes p_2, q_2 , computes $n_2 = p_2 * q_2$ and then publishes n_2 . Besides, FA and HA share a secret key K_{FH} .

2.1 Registration Phase

This phase is involved if a mobile user MU desires to access the system. MU selects a password PW_M and a random number b_M and computes $h(PW_M \parallel b_M)$. He then sends his identity ID_M and $h(PW_M \parallel b_M)$ to HA in a secure channel. Once HA receives the message from MU , he computes

$$u = (h(ID_M \parallel d))^2 \bmod n_1,$$

$$C = h(u \parallel ID_M),$$

$$v = u \oplus h(PW_M \parallel b_M).$$

Then, *HA* issues a smart card contains $\{n_1, h(\cdot), v, C\}$ and sends it to *MU* through a secure channel. Finally, *MU* stores the random number b_M into the smart card (Fig. 1).

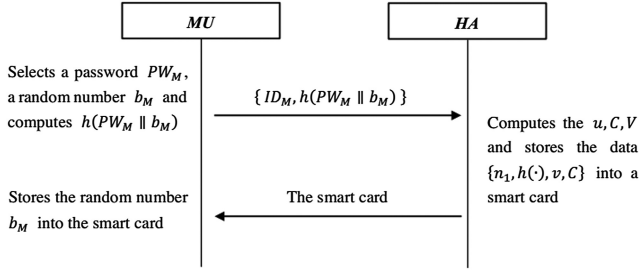


Fig. 1. Registration phase of Liu *et al.*'s protocol.

2.2 Mutual Authentication Phase

In this phase, as shown in Fig. 2, *MU* and *FA* perform the mutual authentication and establish a session key under the assistance of *HA*.

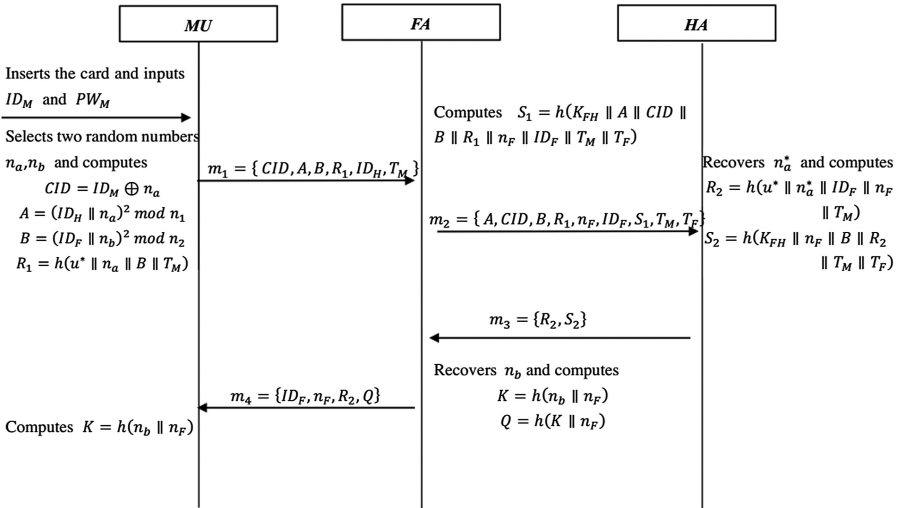


Fig. 2. The authentication and key agreement phase of Liu *et al.*'s protocol.

Step 1. *MU* inserts his smart card into the smart card reader and inputs his ID_M and PW_M . The device then calculates $u^* = v \oplus h(PW_M \parallel b_M)$, $C^* = h(u^* \parallel ID_M)$. Next, the device will check $C^*? = C$. If the value of C^* is not equal to that of C , the device terminates this login request for a period of time. Otherwise, the device randomly selects n_a, n_b and computes $CID = ID_M \oplus n_a$, $A = (ID_H \parallel n_a)^2 \bmod n_1$, $B = (ID_F \parallel n_b)^2 \bmod n_2$ and $R_1 = h(u^* \parallel n_a \parallel B \parallel T_M)$. Then, it sends the message $m_1 = \{CID, A, B, R_1, ID_H, T_M\}$ to *FA*.

Step 2. Once *FA* receives the message m_1 , *FA* checks whether the timestamp T_M is valid. If it is valid, *FA* selects a random number n_F and continues to compute $S_1 = h(K_{FH} \parallel A \parallel CID \parallel B \parallel R_1 \parallel n_F \parallel ID_F \parallel T_M \parallel T_F)$. Then it sends the message $m_2 = \{A, CID, B, R_1, n_F, ID_F, S_1, T_M, T_F\}$ to *HA*.

Step 3. Upon receiving the message m_2 from *FA*, *HA* first checks whether the timestamp T_F equals the current time or not. If so, *HA* then computes $S_1^* = h(K_{FH} \parallel A \parallel CID \parallel B \parallel R_1 \parallel n_F \parallel ID_F \parallel T_M \parallel T_F)$ and checks $S_1^*? = S_1$. If the two values are equal, *HA* obtains n_a^* from *MSR* of *A* with the knowledge of ID_H , and computes $ID_M^* = CID \oplus n_a^*$, $u^* = (h(ID_M^* \parallel d))^2 \bmod n_1$, $R_1^* = h(u^* \parallel n_a^* \parallel B \parallel T_M)$. Next, *HA* checks $R_1^*? = R_1$. If the equation holds, *HA* computes $R_2 = h(u^* \parallel n_a^* \parallel ID_F \parallel n_F \parallel T_M)$ and $S_2 = h(K_{FH} \parallel n_F \parallel B \parallel R_2 \parallel T_M \parallel T_F)$, then sends the message $m_3 = \{R_2, S_2\}$ to *FA*.

Step 4. After receiving m_3 from *HA*, *FA* computes $S_2^* = h(K_{FH} \parallel n_F \parallel B \parallel R_2 \parallel T_M \parallel T_F)$ and checks whether the equation $S_2^*? = S_2$ holds. If it holds, it means that *MU* is authorized. Then, *FA* obtains n_b from *MSR* of *B* with the knowledge of ID_F . Next, *FA* computes the session key $K = h(n_b \parallel n_F)$, $Q = h(K \parallel n_F)$, and sends the message $m_4 = \{ID_F, n_F, R_2, Q\}$ to *MU*.

Step 5. Upon receiving m_4 from *FA*, *MU* computes $R_2^* = h(u^* \parallel n_a^* \parallel ID_F \parallel n_F \parallel T_M)$ and checks whether the equation $R_2^*? = R_2$ holds. If it holds, it means that *FA* is authorized. Then, *MU* computes $K = h(n_b \parallel n_F)$, $Q^* = h(K \parallel n_F)$ and verifies $Q^*? = Q$. If so, *MU* believes that K is the session key between him and *FA*.

2.3 Updated Password Phase

When *MU* wants to change his password, he inserts the smart card into the device and input his ID_M and PW_M . The device computes

$$u = v \oplus h(PW_M \parallel b_M),$$

$$C^* = h(u \parallel ID_M).$$

Then, the device checks $C^*? = C$. If the value of C^* is not equal to that of C , the device terminates the password change phase for a period of time. Otherwise, *MU* inputs a new password PW'_M and the device computes v' to replace v on the memory of the smart card.

$$v' = v \oplus h(PW_M \parallel b_M) \oplus h(PW'_M \parallel b_M)$$

3 Cryptanalysis of Liu et al.'s Protocol

For the past decade, various authenticated key exchange protocols have been proposed but many of them have been proven insecure [9–13]. In Liu *et al.*'s protocol, they claimed that their protocol can provide user anonymity and resist offline password guessing attack even when the smart card is disclosed. However, we find that this protocol still fails to protect user anonymity. More specifically, with the message eavesdropped from the authentication phase, an adversary can obtain the identity of the *MU*. Besides, if the secret information in the corresponding card is exposed, the adversary can easily carry out an offline password guessing attack.

3.1 Fails to Protect User Anonymity

The anonymity protects the privacy of participating entities, which ensures the untraceability of them during communication. The most common way is to conceal user's real identity. In Liu *et al.*'s paper, *XOR* operation and random number are used to protect the user anonymity. Unfortunately, we find that if an adversary *E* eavesdrops the login message m_1 , he can guess the *MU*'s identity in polynomial time by the following steps.

Step 1. The adversary *E* eavesdrops the public channel and obtains the message m_1 , he then extracts R_1, A, CID from m_1 .

Step 2. *E* guesses the value ID'_M and computes $n'_a = CID \oplus ID'_M, A' = (ID'_M \parallel n'_a)^2 \bmod n_1$. Then verifies $A' ? = A$.

Step 3. If the verification succeeds, the adversary considers ID'_M as the *MU*'s identity. Otherwise, he repeats *Step 1*.

Since the identity guessing attack mentioned above does not need to interact with *FA* and *HA*, it is easy to launch and succeed in polynomial time. Once the adversary guesses the *MU*'s identity, he can trace the message from it. Thus, Liu *et al.*'s cannot provide user anonymity.

3.2 Suffer from the Offline Password Guessing Attack

From the aforementioned analysis, the adversary can obtain the *MU*'s identity by performing offline identity guessing attack. Then, we demonstrate Liu *et al.*'s protocol suffers from password offline guessing attack using the compromised identity ID_M and the information stored in the smart card. The adversary can guess the password as follows:

Step 1. The adversary *E* eavesdrops the public channel and obtains the message m_1 , he then extracts R_1, CID, T_M from m_1 .

Step 2. The adversary guesses the password PW'_M and computes $u' = v \oplus h(PW'_M \parallel b_M), n_a = CID \oplus ID_M, R'_1 = h(u' \parallel n_a \parallel B \parallel T_M)$, where b_M and v are stored in the card and ID_M is gained from offline identity guessing attack. Then verifies $R'_1 ? = R_1$.

Step 3. If the two values equal, the adversary believes PW'_M is MU 's password. Otherwise, he repeats *Step 1*.

Since the MU 's password can be guessed by the value of R_1 , Liu *et al.*'s cannot withstand offline password guessing attack.

4 Possible Improvement

The reason for such attacks is because the adversary can successfully obtain the identity of the MU . To overcome the weakness, a simple but effective solution is to use one-way hash function to protect the random number n_a . In the *step 1* of the authentication process, we can modify the calculation of CID , using $CID = ID_M \oplus h(n_a)$ instead.

Theorem 1. Our patched protocol has the property of user anonymity.

Correctness. In Liu *et al.*'s protocol, the adversary can guess the identity of MU , and then successfully compute the parameter A' and verify the equality $A' = A$ holds or not. In the patched protocol, with the message $m_1 = \{CID, A, B, R_1, ID_H, T_M\}$ acquired in the login phase, it is very hard for an adversary to derive MU 's identity ID_M from CID and A , where $CID = ID_M \oplus h(n_a)$, and $A = (ID_H \parallel n_a)^2 \bmod n_1$. To carry out an off-line identity guessing attack, let the adversary E guesses the MU 's identity ID'_M . If the adversary wants to verify the identity, he has to obtain the value of n_a and A' , where $n'_a = CID \oplus ID'_M$, $A' = (ID'_M \parallel n'_a)^2 \bmod n_1$. However, the value he can only get is $h(n_a)$. As $h(\cdot)$ is one-way hash function, extract n_a from $h(n_a)$ is almost impossible. He cannot gain the value of A' due to the difficulty of extracting MSR of a quadratic residue modulo n . Hence, our patched protocol can provide user anonymity.

Theorem 2. Our patched protocol can withstand an off-line password guessing attack.

Correctness. In Liu *et al.*'s protocol, the adversary can obtain the identity of MU and steal the corresponding smart card. With the data $\{n_1, h(\cdot), v, C, b_M\}$ extracted from the smart card, the adversary can guess the possible password and compute the parameter R'_1 and use the equality $R'_1 = R_1$ to verify the result. In our patched protocol, let the adversary guess the possible password PW'_M . To verify the value, he needs to compute u', n_a and R'_1 , where $u' = v \oplus h(PW'_M \parallel b_M)$, $h(n_a) = CID \oplus ID_M$, $R'_1 = h(u' \parallel n_a \parallel B \parallel T_M)$. It is clear with the protection of $h(\cdot)$, n_a cannot be obtained. Hence, the value of R'_1 cannot be computed. Thus, the adversary has no ways to judge whether the PW'_M is correct or not. In a word, our patched protocol can resist an off-line password guessing attack.

5 Conclusions

In this paper, we analyze the remote user authentication protocol with smart cards for wireless communication proposed by Liu *et al.* Although their protocol uses *MSR* to ensure the security, it is still subjected to two security issues, the failure of user anonymity and password protection. We later provide a simple but effective patch to enhance the protocol so that it can resist the two security issues.

Acknowledgments. The work of Chien-Ming Chen was supported in part by the Project NSFC (National Natural Science Foundation of China) under Grant number 61402135 and in part by Shenzhen Technical Project under Grant number JCYJ20170307151750788. The work of Eric Ke Wang was supported in part by National Natural Science Foundation of China (No. 61572157), grant No. 2016A030313660 from Guangdong Province Natural Science Foundation, JCYJ2016 0608161351559 from Shenzhen Municipal Science and Technology Innovation Project.

References

1. Lee, C.C., Hwang, M.S., Liao, I.E.: Security enhancement on a new authentication scheme with anonymity for wireless environments. *IEEE Trans. Ind. Electron.* **53**(5), 1683–1687 (2006)
2. Jing, X., Zhu, W.T., Feng, D.G.: An efficient mutual authentication and key agreement protocol preserving user anonymity in mobile networks. *Comput. Commun.* **34**(3), 319–325 (2011)
3. Wang, X., Zhao, J.: An improved key agreement protocol based on chaos. *Commun. Nonlinear Sci. Numer. Simul.* **15**(12), 4052–4057 (2010)
4. Guo, C., Chang, C.C., Sun, C.Y.: Chaotic maps-based mutual authentication and key agreement using smart cards for wireless communications. *J. Inf. Hiding Multimedia Sig. Process.* **4**(2), 99–109 (2013)
5. Liu, C.-S., et al.: Mutual authentication with anonymity for roaming service with smart cards in wireless communications. In: *International Conference on Network and System Security*. Springer (2016)
6. Jebek, E.: Integer factoring and modular square roots. *J. Comput. Syst. Sci.* **82**, 380–394 (2016)
7. Rabin, M.O.: Digitalized signatures and public-key functions as intractable as factorization. Technical report, Cambridge, MA, USA (1979)
8. Williams, H.C.: A modification of the RSA public-key encryption procedure (cor-resp.). *IEEE Trans. Inf. Theor.* **26**(6), 726–729 (1980)
9. Chen, C.M., Li, C.M., Liu, S., Wu, T.Y., Pan, J.S.: A provable secure private data delegation scheme for mountaineering events in emergency system. *IEEE Access* **5**, 3410–3422 (2017)
10. Chen, C.M., Fang, W., Wang, K.H., Wu, T.Y.: Comments on an improved secure and efficient password and chaos-based two party key agreement protocol. *Nonlinear Dyn.* **87**(3), 2073–2075 (2017)
11. Chen, C.M., Xu, L., Wu, T.Y., Li, C.R.: On the security of a chaotic maps-based three-party authenticated key agreement protocol. *J. Netw. Intell.* **1**(2), 61–65 (2016)

12. Sun, H.M., He, B.Z., Chen, C.M., Wu, T.Y., Lin, C.H., Wang, H.: A provable authenticated group key agreement protocol for mobile environment. *Inf. Sci.* **321**, 224–237 (2015)
13. Chen, C.M., Wang, K.H., Wu, T.Y., Pan, I.S., Sun, H.M.: A scalable transitive human-verifiable authentication protocol for mobile devices. *IEEE Trans. Inf. Forensics Secur.* **8**(8), 1318–1330 (2013)

Comments on Islam Et Al.'s Certificateless Designated Server Based Public Key Encryption with Keyword Search Scheme

Tsu-Yang Wu^{1,2(✉)}, Chao Meng³, King-Hang Wang⁴, Chien-Ming Chen³,
and Jeng-Shyang Pan^{1,2}

¹ Fujian Provincial Key Lab of Big Data Mining and Applications,
Fujian University of Technology, Fuzhou 350118, China
wutsuyang@gmail.com, jengshyangpan@fjut.edu.cn

² National Demonstration Center for Experimental Electronic Information and
Electrical Technology Education, Fujian University of Technology,
Fuzhou 350118, China

³ School of Computer Science and Technology, Harbin Institute of Technology -
Shenzhen, Shenzhen 518055, China
171521532@qq.com, chienming.taiwan@gmail.com

⁴ Department of Computer Science and Engineering, Hong Kong University
of Science and Technology, Clear Water Bay, Kowloon, Hong Kong
kevinw@cse.ust.hk

Abstract. Recently, Islam et al. proposed a certificateless designated server based public key encryption with keyword search (CL-dPEKS) scheme which combines the concepts of dPEKS and certificateless public key cryptosystem. In this paper, we show that their scheme does not provide the ciphertext and the trapdoor indistinguishabilities, two important security notions of dPEKS. Concretely, we demonstrate that their CL-dPEKS scheme suffered from off-line keyword guessing attacks on ciphertext and trapdoor by outside adversary and malicious server.

Keywords: Public key encryption with keyword search · Designated server · Certificateless · Off-line keyword guessing attacks · Cryptanalysis

1 Introduction

With the fast growth of cloud technologies [13], enterprises or people can choose to outsource their data to cloud. However, it lacks the physical control of their data by their own such that the cloud may try to know the outsourced data. To solve this security risk, the outsourced data should be encrypted before outsourcing to the cloud. However, it is occurred how to retrieve an encrypted data in cloud.

Public key encryption with keyword search (PEKS) (or called searchable encryption) [3] is a cryptographic primitive introduced by Boneh et al. in 2004.

The flowchart of Boneh et al.’s PEKS scheme is depicted in Fig. 1. There are three roles existed in the environment: a data owner, a server, and a data user. Note that the data user who can be the data owner himself or any other designated individual who has the right of accessing the data. The flowchart of PEKS scheme is described as follows.

1. The data owner first encrypts the keywords with the data user’s public key and uploaded to the server together with the encrypted data files.
2. When a data user wish to retrieve data file with a particular keyword, she/he will generate a trapdoor using own private key and the keyword that wants to search. This trapdoor is securely sent to the server.
3. The server can test an encrypted keyword ciphertext whether matching with the trapdoor using some mathematical computations. If true, the matching encrypted data will then sent to the user.

Such framework was used in the subsequent works such as supporting conjunctive keywords [15] and supporting multi users [8].

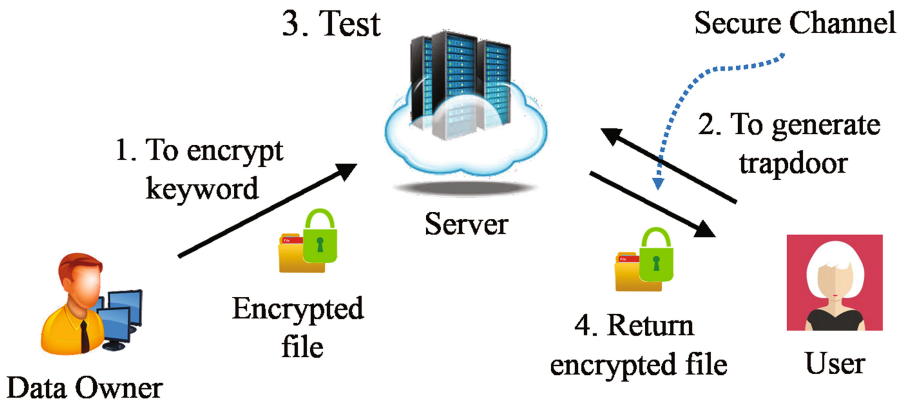


Fig. 1. The flowchart of Boneh et al.’s PEKS scheme [3]

In order to remove the required secure channel in Boneh et al.’s PEKS model, Baek et al. [2] redefined a new model and proposed a new scheme called PEKS with designated server (dPEKS). However, Rhee et al. [17] pointed out that Baek et al.’s model seriously limits the ability of the adversary. They also enhanced the security model of the dPEKS scheme. Later, Rhee et al. [18] defined a new security notion of dPEKS called “Trapdoor indistinguishability” which allows a scheme to be formally proven secure against a non-designated person who wants to launch an off-line keyword guessing attack [5, 7, 11, 19, 20, 23]. Note that in the kinds of attacks, an outside attacker or an inside attacker (malicious server) can simply enumerate on all possible keywords to construct an encrypted meta-data and test that with the ciphertext, the trapdoor, or both.

In 2001, Boneh and Franklin first developed identity (ID)-based public key cryptosystems [4, 21] to solve the certificate managements in the traditional public key cryptosystems so that one may use the ID of a person as its public key, where the public key owner, has the corresponding private key generated by a trusted private key generator (PKG). However it causes a key escrow problem where the PKG can actually derive any users' private key in this system and it could lead to a complete break down of the entire cryptosystem if the master key of the PKG is leaked or misused. In order to solve this problem, Al-Riyami and Paterson [1] first introduced the concept of certificateless public key cryptosystem. In their cryptosystem, PKG only computes the user U 's partial private key with his identifier ID_U and the PKG's master key. U then combines its partial private key with some secret information (selected by U) to generate its actual private key. Meanwhile, the U also combines its secret information with the PKG's public parameters to compute its public key. In particular, there is no certificate for U 's public key.

Recently, Islam et al. [9] combined the concepts of dPEKS and certificateless public key cryptosystem to propose a certificateless designated server based public key encryption with keyword search (or called CL-dPEKS for short) scheme. In this paper, we demonstrate that their CL-dPEKS scheme also insecure against off-line keyword guessing attacks on ciphertext and trapdoor by outside adversary and malicious server. In other words, both attackers can distinguish ciphertext and trapdoor with the guessed keyword successfully. It seriously violates the security notions of dPEKS.

2 Review of Islam Et Al.'s CL-dPEKS Scheme

2.1 Bilinear Pairing

Let $\mathbb{E}(\mathbb{F}_p)$ be an elliptic curve over a finite field \mathbb{F}_p . Then, we can obtain the collection $\{(x, y) | (x, y) \in \mathbb{E}(\mathbb{F}_p)\}$ with an infinite point O under addition operation "+" formed an abelian group $\mathbb{E}_{\mathbb{F}_p}(x, y)$.

To select \mathbb{G}_1 be an additive cyclic subgroup of $\mathbb{E}(\mathbb{F}_p)$ and \mathbb{G}_2 be a multiplicative cyclic group over \mathbb{F}_p . A bilinear pairing e is a map (also called pairing or bilinear map) defined by $e : \mathbb{G}_1 \times \mathbb{G}_1 \rightarrow \mathbb{G}_2$ and satisfies the following three properties.

1. Bilinear. For all $P, Q \in \mathbb{G}_1$ and $a, b \in \mathbb{Z}_p^*$, we have $e(aP, bQ) = e(P, Q)^{ab}$.
2. Non-degenerate. For identity $1_{\mathbb{G}_1} \in \mathbb{G}_1$, we have $e(1_{\mathbb{G}_1}, 1_{\mathbb{G}_1})$ also an identity of \mathbb{G}_2 .
3. Computable. For all $P, Q \in \mathbb{G}_1$, there exist several algorithms to compute $e(P, Q)$, for example, Miller's Weil pairing algorithm [14].

For the details about bilinear pairings, readers can refer to [4, 6, 10, 12, 22, 24] for full descriptions.

2.2 Islam Et Al.'s CL-dPEKS Scheme

In this section, we briefly review Islam et al.'s CL-dPEKS scheme. Their scheme consists of eight algorithms: Setup, Gen-Secret-Key, Gen-Partial-Private-Key, Set-Private-Key, Set-Public-Key, Encrypt, Gen-Trapdoor, and Test-Trapdoor described as follows.

1. *Setup*. The PKG inputs a security parameter 1^k and generates following parameters with this algorithm.
 - (a) Selecting a large prime p with size k .
 - (b) Selecting a bilinear map group system $\{\mathbb{F}_p, \mathbb{E}(\mathbb{F}_p), e, \mathbb{G}_1, \mathbb{G}_2, P\}$, where \mathbb{F}_p is a finite field, $\mathbb{E}(\mathbb{F}_p)$ is an elliptic curve over \mathbb{F}_p , e is a bilinear map ($e : \mathbb{G}_1 \times \mathbb{G}_1 \rightarrow \mathbb{G}_2$), P is a generator of \mathbb{G}_1 .
 - (c) Choosing $s \in_R \mathbb{Z}_p^*$ as the master key and the corresponding public key P_{pub} is computed by $P_{pub} = s \cdot P$.
 - (d) Choosing two cryptographic hash functions $H : \{0, 1\}^* \rightarrow \mathbb{G}_1$ and $h : \{0, 1\}^* \rightarrow \{0, 1\}^l$.

Finally, the PKG publishes public parameters

$$param = \{\mathbb{F}_p, \mathbb{E}(\mathbb{F}_p), e, \mathbb{G}_1, \mathbb{G}_2, p, P, P_{pub}, H, h\}.$$

2. *Gen-Secret-Key*. Entity E_i with identity ID_i selects $x_i \in_R \mathbb{Z}_p^*$ as secret key and then computes the corresponding public key $P_i = x_i \cdot P$ for $i = \{\text{client}(C), \text{cloud server}(S)\}$.
3. *Gen-Partial-Private-Key*. Upon receiving (ID_i, P_i) from E_i , the PKG runs this algorithm to generate E_i 's partial private key and partial public key as follows.
 - (a) Selecting $t_i \in_R \mathbb{Z}_p^*$.
 - (b) Computing $T_i = t_i \cdot P$, $l_i = h(ID_i, T_i, P_i)$, and $d_i = (t_i + s \cdot l_i) \bmod p$.
Finally, the PKG sends (d_i, T_i) to E_i via a secure channel. The partial private key of E_i with ID_i is defined by d_i and the partial pulic key of E_i with ID_i is defined by $d_i \cdot P$.
4. *Set-Private-Key*. Entity E_i sets its private key sk_i as $sk_i = (x_i, d_i)$ for $i \in \{C, S\}$.
5. *Set-Public-Key*. Entity E_i sets its public key pk_i as $pk_i = (P_i, T_i)$ for $i \in \{C, S\}$.
6. *Encrypt*. To encrypt a keyword set $\mathcal{W} = \{w_1, w_2, \dots, w_n\}$, client C inputs its identity ID_C , its public key (P_C, T_C) , server's identity ID_S , server's public key (P_S, T_S) , and \mathcal{W} and then runs this algorithm as follows.
 - (a) Selecting $r \in_R \mathbb{Z}_p^*$ and computing $U = r \cdot P$.
 - (b) Computing $z_i = r \cdot [H(w_i) + P_S + T_S + l_S \cdot P_{pub}]$ for $i = 1, 2, \dots, n$, where $l_S = h(ID_S, T_S, P_S)$.

The ciphertext of \mathcal{W} is defined by $\mathcal{C}_{\mathcal{W}} = (U, z_1, z_2, \dots, z_n)$.

7. *Gen-Trapdoor*. To generate a trapdoor T_{w_j} of keyword w_j , client C inputs its identity ID_C , its private key $sk_C = (x_C, d_C)$, and w_j and then computes $T_{w_j} = (x_C + d_C) \cdot H(w_j)$.

8. *Test-Trapdoor*. To teste ciphertext $\mathcal{C}_{\mathcal{W}} = (U, z_1, z_2, \dots, z_n)$ whether containing keyword w_j , cloud server S inputs $\mathcal{C}_{\mathcal{W}}$, its private key (x_S, d_S) , client C 's public key (P_C, T_C) , and trapdoor T_{w_j} and then verifies

$$e(T_{w_j} + (x_S + d_S) \cdot (P_C + T_C + l_C \cdot P_{pub}), U) \stackrel{?}{=} e(z_i, P_C + T_C + l_C \cdot P_{pub}),$$

where $l_C = h(ID_C, T_C, P_C)$. If the verification is true, it means that the keyword w_j in T_{w_j} is contained in the keyword set \mathcal{W} . Otherwise, S terminates the process.

3 Cryptanalysis of Islam Et Al.'s CL-dPEKS Scheme

In this section, we demonstrate an outside adversary can launch keyword guessing attacks on ciphertext and trapdoor in Islam et al.'s CL-dPEKS scheme. It is easy to see that if the cloud server S is honest but curious, then S can also launch keyword guessing attacks on ciphertext and trapdoor with the same attack procedures.

3.1 Keyword Guessing Attack on Ciphertext

Assume that an outside attacker \mathcal{A} captures $\mathcal{C}_{\mathcal{W}} = (U, z_1, z_2, \dots, z_n)$, the ciphertext of keyword set \mathcal{W} . Then, \mathcal{A} can launch an off-line keyword guessing attack to test Z_i contains which keyword as follows.

1. To compute $\Theta = P_S + T_S + l_S \cdot P_{pub}$, where $l_S = h(ID_S, P_S, T_S)$, ID_S is the server S 's idenity, and (P_S, T_S) is the S 's public key.
2. To guess an appropriate keyword w' .
3. To verify

$$e(z_i, P) \stackrel{?}{=} e(H(w') + \Theta, U) \text{ for } i = 1, 2, \dots, n.$$

If the verification is true, it means that z_i is generated by w' . Otherwise, \mathcal{A} goes back to the step 2 and continues to execute the step 3.

Here, we present the correctness of our attack. Assume that w' is the success guessed keyword. Then,

$$e(z_i, P) = e(r \cdot [H(w_i) + P_S + T_S + l_S \cdot P_{pub}], P) = e(H(w_i) + \Theta, U),$$

where $U = r \cdot P$.

3.2 Keyword Guessing Attack on Trapdoor

Assume that an outside attacker \mathcal{A} captures $T_{w_j} = (x_C + d_C) \cdot H(w_j)$, the trapdoor of keyword w_j . Then, \mathcal{A} can launch an off-line keyword guessing attack to test T_{w_j} contains which keyword as follows.

1. To compute $\Lambda = P_C + T_C + l_C \cdot P_{pub}$, where $l_C = h(ID_C, P_C, T_C)$, ID_C is the client C 's idenity, and (P_C, T_C) is the C 's public key.

2. To guess an appropriate keyword w' .
3. To verify

$$e(T_{w_j}, P) \stackrel{?}{=} e(H(w'), A).$$

If the verification is true, it means that T_{w_j} is generated by w' . Otherwise, \mathcal{A} goes back to the step 2 and continues to execute the step 3.

Here, we present the correctness of our attack. Assume that w' is the success guessed keyword. Then,

$$e(T_{w_j}, P) = e((x_C + d_C) \cdot H(w_j), P) = e(H(w_j), (x_C + d_C) \cdot P) = e(H(w_j), A),$$

where

$$P_C + T_C + l_C \cdot P_{pub} = (x_C + t_S + l_C \cdot s) \cdot P = (x_C + d_C) \cdot P.$$

4 Conclusion

In this paper, we have shown that Islam et al.'s CL-dPEKS scheme does not provide the ciphertext and the trapdoor indistinguishabilities. Up to now, two certificateless based dPEKS schemes were proposed in [16, 25]. Unfortunately, they seem insecure against off-line keyword guessing attacks. To design a new security model and secure scheme to overcome the known attacks are valuable works in the future.

Acknowledgments. The authors would thank anonymous referees for a valuable comments and suggestions. The work of Chien-Ming Chen was supported in part by the Project NSFC (National Natural Science Foundation of China) under Grant number 61402135 and in part by Shenzhen Technical Project under Grant number JCYJ20170307151750788.

References

1. Al-Riyami, S.S., Paterson, K.G.: Certificateless public key cryptography. In: Advances in Cryptology-ASIACRYPT 2003, pp. 452–473. Springer (2003)
2. Baek, J., Safavi-Naini, R., Susilo, W.: Public key encryption with keyword search revisited. In: Computational Science and Its Applications-ICCSA 2008, pp. 1249–1259 (2008)
3. Boneh, D., Di Crescenzo, G., Ostrovsky, R., Persiano, G.: Public key encryption with keyword search. In: Advances in Cryptology-Eurocrypt 2004, pp. 506–522. Springer (2004)
4. Boneh, D., Franklin, M.: Identity-based encryption from the weil pairing. In: Annual International Cryptology Conference, pp. 213–229. Springer (2001)
5. Byun, J.W., Rhee, H.S., Park, H.A., Lee, D.H.: Off-line keyword guessing attacks on recent keyword search schemes over encrypted data. In: Workshop on Secure Data Management, pp. 75–83. Springer (2006)
6. Chen, L., Cheng, Z., Smart, N.P.: Identity-based key agreement protocols from pairings. Int. J. Inf. Secur. **6**(4), 213–241 (2007)

7. Hu, C., Liu, P.: An enhanced searchable public key encryption scheme with a designated tester and its extensions. *J. Comput.* **7**(3), 716–723 (2012)
8. Hwang, Y., Lee, P.: Public key encryption with conjunctive keyword search and its extension to a multi-user system. In: *Pairing-Based Cryptography-Pairing 2007*, pp. 2–22 (2007)
9. Islam, S.H., Obaidat, M.S., Rajeev, V., Amin, R.: Design of a certificateless designated server based searchable public key encryption scheme. In: *International Conference on Mathematics and Computing*, pp. 3–15. Springer (2017)
10. Li, C.T., Wu, T.Y., Chen, C.L., Lee, C.C., Chen, C.M.: An efficient user authentication and user anonymity scheme with provably security for iot-based medical care system. *Sensors* **17**(7), 1482 (2017)
11. Lu, Y., Wang, G., Li, J., Shen, J.: Efficient designated server identity-based encryption with conjunctive keyword search. *Ann. Telecommun.* **72**(5–6), 359–370 (2017)
12. Ma, H., Zhang, Z., Li, H., Yin, S.L., Chu, Z.: A provable private data aggregation scheme based on digital signatures and homomorphic encryption for wireless sensor networks. *J. Inf. Hiding Multimedia Signal Process.* **8**(3), 536–543 (2017)
13. Mell, P., Grance, T., et al.: *The NIST definition of cloud computing* (2011)
14. Miller, V.S.: The weil pairing, and its efficient calculation. *J. Cryptol.* **17**(4), 235–261 (2004)
15. Park, D.J., Kim, K., Lee, P.J.: Public key encryption with conjunctive field keyword search. In: *International Workshop on Information Security Applications*, pp. 73–86. Springer (2004)
16. Peng, Y., Cui, J., Peng, C., Ying, Z.: Certificateless public key encryption with keyword search. *Chin. Commun.* **11**(11), 100–113 (2014)
17. Rhee, H.S., Park, J.H., Susilo, W., Lee, D.H.: Improved searchable public key encryption with designated tester. In: *Proceedings of the 4th International Symposium on Information, Computer, and Communications Security*, pp. 376–379. ACM (2009)
18. Rhee, H.S., Park, J.H., Susilo, W., Lee, D.H.: Trapdoor security in a searchable public-key encryption scheme with a designated tester. *J. Syst. Soft.* **83**(5), 763–771 (2010)
19. Rhee, H.S., Susilo, W., Kim, H.J.: Secure searchable public key encryption scheme against keyword guessing attacks. *IEICE Electron. Express* **6**(5), 237–243 (2009)
20. Wang, B., Chen, T., Jeng, F.: Security improvement against malicious server's attack for a dpeks scheme. *Int. J. Inf. Edu. Technol.* **1**(4), 350 (2011)
21. Wu, T.Y., Tsai, T.T., Tseng, Y.M.: Efficient searchable ID-based encryption with a designated server. *Annals of telecommunications-Annales des télécommunications* **69**(7–8), 391–402 (2014)
22. Wu, T.Y., Tseng, Y.M.: An ID-based mutual authentication and key exchange protocol for low-power mobile devices. *Comput. J.* **53**(7), 1062–1070 (2010)
23. Yau, W.C., Phan, R.C.W., Heng, S.H., Goi, B.M.: Keyword guessing attacks on secure searchable public key encryption schemes with a designated tester. *Int. J. Comput. Math.* **90**(12), 2581–2587 (2013)
24. Yin, S.L., Li, H., Liu, J.: A new provable secure certificateless aggregate signcryption scheme. *J. Inf. Hiding Multimedia Signal Process.* **7**(6), 1274–1281 (2016)
25. Zheng, Q., Li, X., Azgin, A.: Clks: certificateless keyword search on encrypted data. In: *International Conference on Network and System Security*, pp. 239–253. Springer (2015)

Author Index

A

Aung, Htin Lynn, [173](#), [182](#)

B

Belhadi, Asma, [59](#)

C

Chang, Chin-Chen, [101](#)

Chen, Bor-An, [26](#)

Chen, Chien-Ming, [191](#), [199](#)

Chen, Chin-Ling, [153](#)

Chen, Chun-Hao, [20](#)

Chen, Jun-Hong, [12](#)

Chen, Yi-Chung, [12](#)

Chen, Yue-Xun, [153](#)

Chiang, Ming-Chao, [37](#)

Chiang, Tsung-Che, [3](#)

Chung, Yi-Nung, [77](#)

D

Djenouri, Youcef, [59](#)

F

Fan, Chia-Chen, [121](#)

Fournier-Viger, Philippe, [59](#), [66](#)

H

Hama, Hiromitsu, [52](#), [165](#)

Hong, Tzung-Pei, [37](#), [66](#)

Hsu, Chao-Hsing, [77](#)

Hu, Yun-Jhong, [77](#)

Huang, Wei-Ming, [37](#)

Huang, Yen-Chia, [109](#)

Huang, Yung-Fa, [26](#)

L

Lai, Chien-Wen, [77](#)

Lai, Jean, [44](#), [93](#), [128](#)

Lan, Guo-Cheng, [37](#)

Lee, Chen-Yu, [3](#)

Lee, Chin-Feng, [101](#), [109](#)

Lee, Ru-Kam, [12](#)

Lee, Tsair-Fwu, [84](#)

Li, Xiaoyin, [93](#)

Lin, Dong-Peng, [109](#)

Lin, Jerry Chun-Wei, [37](#), [59](#), [66](#)

Lin, Yu-Chuan, [121](#)

M

Meng, Chao, [199](#)

Myroniv, Bohdan, [136](#)

P

Pan, Jeng-Shyang, [144](#), [199](#)

Peng, Zhengdao, [128](#)

Pham, Duc-Tinh, [84](#)

Q

Quan, Li, [44](#)

R

Ren, Yi, [136](#)

S

Shao, Yinan, [66](#)

T

Tan, Tan-Hsu, [26](#)

Tang, Yung-Wen, [153](#)

Tin, Pyke, [52](#), [165](#)

Tran, Chi-Kien, [84](#)

Tsai, Xian-Zhi, [77](#)

Tseng, Chin-Dar, [84](#)

Tseng, Yen-Ming, [144](#)

Tseng, Yu-Chee, [136](#)

W

Wang, Eric Ke, [191](#)
Wang, Guangjie, [191](#)
Wang, King-Hang, [199](#)
Weng, Chi-Yao, [109](#)
Wu, Cheng-Wei, [136](#)
Wu, Tsu-Yang, [66](#), [191](#), [199](#)

X

Xiang, Bin, [191](#)

Y

Yang, Yi-Che, [12](#)
Yeh, Jia-Fong, [3](#)
Yu, Chih-Hung, [20](#)
Yuan, Shyan-Ming, [121](#)

Z

Zeng, Qun-Feng, [101](#)
Zhang, Wenxin, [144](#)
Zhang, Xuebai, [121](#)
Zhou, Yong-Feng, [153](#)
Zin, Thi Thi, [52](#), [165](#), [173](#), [182](#)