Advances in Intelligent Systems and Computing 329

Hui Sun · Ching-Yu Yang
Chun-Wei Lin · Jeng-Shyang Pan
Vaclav Snasel · Ajith Abraham *Editors*

Genetic and Evolutionary Computing

Proceeding of the Eighth International Conference on Genetic and Evolutionary Computing, October 18–20, 2014, Nanchang, China

Springer

Advances in Intelligent Systems and Computing

Volume 329

About this Series

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

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

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

Hui Sun · Ching-Yu Yang
Chun-Wei Lin · Jeng-Shyang Pan
Vaclav Snasel · Ajith Abraham
Editors

Genetic and Evolutionary Computing

Proceeding of the Eighth International Conference on Genetic and Evolutionary Computing, October 18–20, 2014, Nanchang, China

Springer

Editors

Hui Sun
Nanchang Institute of Technology
HIGH-TECH Industry
Jiangxi Province
China

Ching-Yu Yang
Department of Computer Science and Information Engineering
National Penghu University of Science and Technology
Magong City
Taiwan

Chun-Wei Lin
Innovative Information Industry Research Center
Harbin Institute of Technology
Shenzhen University
Shenzhen
China

Jeng-Shyang Pan
Innovative Information Industry Research Center
Harbin Institute of Technology
Shenzhen University
Shenzhen
China

Vaclav Snasel
Faculty of Elec. Eng. and Comp. Sci.
Department of Computer Science
VSB-Technical University of Ostrava
Ostrava
Czech Republic

Ajith Abraham
Machine Intelligence Research Labs (MIR Labs)
Scientific Network for Innovation and Research Excellence
Auburn Washington
USA

ISSN 2194-5357 ISSN 2194-5365 (electronic)
ISBN 978-3-319-12285-4 ISBN 978-3-319-12286-1 (eBook)
DOI 10.1007/978-3-319-12286-1

Library of Congress Control Number: 2014951877

Springer Cham Heidelberg New York Dordrecht London

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

This volume composes the proceedings of the Eighth International Conference on Genetic and Evolutionary Computing (ICGEC 2014), which was hosted by Nanchang Institute of Technology and was held in Nanchang City on October 18-20, 2014. ICGEC 2014 was technically co-sponsored by National Kaohsiung University of Applied Sciences (Taiwan), VSB-Technical University of Ostrava (Czech Republic), and Sping. It aimed to bring together researchers, engineers, and policymakers to discuss the related techniques, to exchange research ideas, and to make friends.

43 excellent papers were accepted for the final proceeding. Three plenary talks were kindly offered by: Ponnuthurai Nagaratnam Suganthan (Nanyang Technological University, Singapore), Chin-Chen Chang (IEEE Fellow, IET Fellow, Feng Chia University, Taiwan), and Bin Hu (IET Fellow, Lanzhou University, China).

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 Nanchang Institute of Technology in making ICGEC 2014 possible.

October 2014

Hui Sun
Ching-Yu Yang
Chun-Wei Lin
Jeng-Shyang Pan
Vaclav Snasel
Ajith Abraham

Conference Organization

Honorary Chair

Zhinong Jin	Nanchang Institute of Technology, China

General Chairs

Hui Sun	Nanchang Institute of Technology, China
Vaclav Snasel	VSB-Technical University of Ostrava, Czech Republic

Program Chairs

Jeng-Shyang Pan	Harbin Institute of Technology Shenzhen Graduate School, China
Ajith Abraham	Machine Intelligence Research Labs (MIR Labs), USA
Ching-Yu Yang	National Penghu University, Taiwan

Technical Chair

Shengqian Wang	Nanchang Institute of Technology, China
	Nanchang Institute of Technology, China

Finance Chair

Peiwu Li	
	Nanchang Institute of Technology, China

Local Chairs

Chengzhi Deng	
Hui Wang	Nanchang Institute of Technology, China
Jia Zhao	Nanchang Institute of Technology, China

Program Committee Members (sorted by last name)

Akira Asano	Hiroshima University, Japan
Mohsen Askari	University of Technology Sydney, Australia
Jonathan Hoyin Chan	King Mongkut's University of Technology Thonburi, Thailand
Chin-Chen Chang	Feng Chia University, Taiwan
Feng-Cheng Chang	Tamkang University, Taiwan
Chien-Ming Chen	Harbin Institute of Technology Shenzhen Graduate School, China
Chao-Chun Chen	Southern Taiwan University, Taiwan
Mu-Song Chen	Da-Yeh University, Taiwan
Pei-Yin Chen	National Cheng Kung University, Taiwan
Shi-Jay Chen	National United University, Taiwan
Yueh-Hong Chen	Far East University, Taiwan
Tsung-Che Chiang	National Taiwan Normal University, Taiwan
Rung-Ching Chen	Chaoyang University of Technology, Taiwan
Shu-Chuan Chu	Cheng-Shiu University, Taiwan
Yi-Nung Chung	National Changhua University of Education, Taiwan
Maurice Clerc	Independent Consultant, French
Jose Alfredo F. Costa	Federal University (UFRN), Brazil
Martine De Cock	Ghent University, Belgium
Zhihua Cui	Taiyuan University of Science and Technology, China
Wang Feng	Kunming University of Science and Technology, China
Xiao-Zhi Gao	Helsinki University of Technology, Finland
Alexander Gelbukh	National Polytechnic Institute, Mexico
Gheorghita Ghinea	Brunel University, UK
Massimo De Gregorio	Istituto di Cibernetica, Italia
Stefanos Gritzalis	University of the Aegean, Greece
Ramin Halavati	Sharif University of Technology, Iran
Chian C. Ho	National Yunlin University of Science Technology, Taiwan
Jiun-Huei Ho	Cheng Shiu University, Taiwan
Cheng-Hsiung Hsieh	Chaoyang University of Technology, Taiwan
Wu-Chih Hu	National Penghu University of Science and Technology, Taiwan
Shu-Hua Hua	Jinwen University of Science and Technology, Taiwan
Hsiang-Cheh Huang	National University of Kaohsiung, Taiwan
Deng-Yuan Huang	Dayeh University, Taiwan

Tien-Tsai Huang	Lunghwa University of Science and Technology, Taiwan
Yung-Fa Huang	Chaoyang University of Technology, Taiwan
Chien-Chang Hsu	Fu-Jen Catholic University, Taiwan
Ming-Wen Hu	Tamkang University, Taiwan
Yongjian Hu	South China University of Technology, China
Donato Impedovo	Politecnico di Bari, Italy
Albert B. Jeng	Jinwen University of Science and Technology, Taiwan
Isabel Jesus	Institute of Engineering of Porto, Portugal
Jie Jing	Zhejiang University of Technology, China
Estevam R. Hruschka Jr.	Federal University of Sao Carlos, Brazil
Mario Koeppen	Kyushu Institute of Technology, Japan
Hsu-Yang Kung	National Pingtung University of Science and Technology, Taiwan
Yau-Hwang Kuo	National Cheng kung University, Taiwan
Kun-Huang Kuo	Chienkuo Technology University, Taiwan
Weng Kin Lai	MIMOS Berhad, Malaysia
Jenn-Kaie Lain	National Yunlin University of Science and Technology, Taiwan
Kwok-Yan Lam	Tsinghua University, China
Sheau-Dong Lang	Central Florida University, USA
Chang-Shing Le	National University of Tainan, Taiwan
Huey-Ming Lee	Chinese Culture University, Taiwan
Shie-Jue Lee	National Sun Yat-Sen University, Taiwan
Jorge Nunez Mc Leod	Universidad Nacional de Cuyo, Argentina
Jung-San Lee	Feng Chia University, Taiwan
Chang-Tsun Li	University of Warwick, UK
Yue Li	Nankai University, China
Jun-Bao Li	Harbin Institute of Technology, China
Guan-Hsiung Liaw	I-Shou University, Taiwan
Chia-Chen Lin	Providence University, Taiwan
Tsung-Chih Lin	Feng-Chia University, Taiwan
Yuh-Chung Lin	Tajen University, Taiwan
Chun-Wei Lin	Harbin Institute of Technology Shenzhen Graduate School, China
Wen-Yang Lin	National University of Kaohsiung, Taiwan
Lily Lin	China University of Technology, Taiwan
Yu-lung Lo	Chaoyang University of Technology, Taiwan
Heitor Silverio Lopes	Federal University of Technology Parana, Brazil
Tianhua Liu	Shenyang Normal University, China
Yuh-Yih Lu	Minghsin University of Science and Technology, Taiwan

Yuchi Ming	Huazhong University of Science and Technology, China
Marco Mussetta	Politecnico Di Torino, Italy
Kazumi Nakamatsu	University of Hyogo, Japan
Julio Cesar Nievola	Pontificia Universidade Catolica do Parana, Brazil
Yusuke Nojima	Osaka Prefecture University, Japan
Isabel L. Nunes	Universidade Nova Lisboa, Portugal
Jae C. Oh	Syracuse University, USA
S.N. Omkar	Indian Institute of Science, Indian
Djamila Ouelhadj	University of Portsmouth, UK
Mauricio Papa	University of Tulsa, USA
Sylvain Piechowiak	University Of Valenciennes, France
Aurora Trinidad Ramirez Pozo	Federal University of Parana, Brazil
Mohammed Al Rashidi	Public Authority for Applied Education and Training (PAAET), Kuwait
Andri Riid	Tallinn University of Technology, Estonia
Selva S. Rivera	Universidad Nacional de Cuyo, Argentina
Lotfi Ben Romdhane	Lotfi Ben Romdhane, Tunisia
A. Fuster-Sabater	Institute of Applied Physics, Spain
G. Sainarayanan	ICT Academy of Tamil Nadu, India
Luciano Sanchez	Oviedo University, Spain
Mariacarla Staffa	Universita' degli Studi di Napoli Federico II, Italy
Yung-Jong Shiah	Kaohsiung Medical University, Taiwan
S.N. Singh	Indian Institute of Technology Kanpur, Indian
Georgios Ch. Sirakoulis	Democritus University of Thrace, Greece
Shing Chiang Tan	Multimedia University, Malaysia
Tsung-Han Tsai	National Central University, Taiwan
Izzettin Temiz	Gazi University, Turkey
Eiji Uchino	Yamaguchi University, Japan
Michael N. Vrahatis	University of Patras, Greece
Brijesh Verma	Central Queensland University, Australia
Sebastian Ventura	University of Cordoba, Spain
Enrique Herrera-Viedma	University of Granada, Spain
Lei Wang	Xi'an University of Technology, China
Lidong Wang	National Computer Network Emergency Response Technical Coordination Center of China
Ling Wang	Tsinghua University, China
Yin Chai Wang	University Malaysia Sarawak, Malaysia
Shiuh-Jeng Wang	Central Police University, Taiwan
K.W. Wong	City University of Hong Kong, Hong Kong
Michal Wozniak	Wroclaw University of Technology, Portugal
Zhijian Wu	Wuhan University, China
Tsu-Yang Wu	Harbin Institute of Technology Shenzhen Graduate School, China

Qingzheng Xu	Xi'an Communication Institute, China
Ruqiang Yan	Southeast University, China
Li Yao	University of Manchester, UK
Chung-Huang Yang	National Kaohsiung Normal University, Taiwan
Chyuan-Huei Thomas Yang	Hsuan Chuang University, Taiwan
Sheng-Yuan Yang	St. John's University, Taiwan
Show-Jane Yen	Mining Chuan University, Taiwan
Fa-xin Yu	Zhejiang University, China
Xinpeng Zhang	Shanghai University, Chian
Yongping Zhang	Hisilicon Technologies Co., Ltd, China
Yong Zhang	Shenzhen University, China
Zhiyong Zhang	Henan University of Science and Technology, China
Xiao-Jun Zeng	University of Manchester, UK
Zhigang Zeng	Huazhong University of Science and Technology, China
Sanyou Zeng	China University of Geosciences (Wuhan), China

Contents

Part I: Nature Inspired Constrained Optimization

Part II: Recent Advances on Evolutionary Optimization Technologies

Part III: Wearable Computing and Intelligent Data Hiding

Part IV: Image Processing and Intelligent Applications

Part V: Intelligent Multimedia Tools and Applications

Part VI: Technologies for Next-Generation Network Environments

Part I

Nature Inspired Constrained Optimization

Artificial Bee Colony Using Opposition-Based Learning

Jia Zhao, Li Lv, and Hui Sun

School of Information Engineering, Nanchang Institute of Technology, No. 289,
Tianxiang Road, State HIGH-Tech Industry Development Zone, Nanchang, 330099, China
zhaojia925@163.com

Abstract. To overcome the drawbacks of artificial bee colony(ABC) algorithm that converges slowly in the process of searching and easily suffers from premature, this paper presents an effective approach, called ABC using opposition-based learning(OBL-ABC). It generates opposite solution by the employed bee and onlooker bee, and chooses the better solution as the new locations of employed bee and onlooker bee according to the greedy selection strategy in order to enlarge the search areas; the new approach proposes a new update rule which can retain the advantages of employed bee and onlooker bee and improve the exploration of OBL-ABC. Experiments are conducted on a set of test functions to verify the performance of OBL-ABC, the results demonstrate promising performance of our method OBL-ABC on convergence and it is suitable for solving the optimization of complex functions.

Keywords: artificial bee colony, opposition-based learning, update rule, optimization.

1 Introduction

Artificial bee clony(ABC), originally developed by Karaboga in 2005, is a swarm intelligence optimization algorithm. Because ABC has many advantages of a few parameters, easy implementation and strong global search optimum, many researchers pay more attention to ABC and apply it into lots of science projects successfully[2,3], such as image processing[4], optimization of electric system[5], parameter estimation [6], economic dispatch[7] and function optimization[8].

Opposition-Based Learning(OBL) , proposed by Tizhoosh[9] in 2005, is successfully applied into genetic algorithm[10], differential evolution algorithm[11], Ant colony algorithm[12] and shuffled frog leaping algorithm[13] to enlarge the search area and enhance the performance of the algorithms.

To improve the convergence velocity of ABC and to avoid the premature convergence, we introduce the opposition-based learning strategy into standard ABC, then present a new method, called artificial bee colony using opposition-based learning (OBL-ABC). In the process of optimization, the algorithm adds opposition-based learning strategy and generates the opposite solution of the employed bee and onlooker bee to enlarge the search area, and chooses the better solution as the new locations of employed bee and onlooker bee according to the greedy selection

H. Sun et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 329, DOI: 10.1007/978-3-319-12286-1_1

strategy; in addition, when the employed bee reaches the search threshold and does not find nectar source, we change it into the scout. Our methods adopt a new nectar source update strategy to update the new location of the scout according to the location of employed bee and to improve the learning ability of the scout and the location exploitation. The simulation results show that OBL-ABC can improve optimization efficiency.

2 ABC Algorithm

There are three kinds of bee, the employed bee, the onlooker bee and the scout. The number of employed bee is equal to the number of the onlooker bee. The employed bee conducts a global search, then all employed bees finish the search, they return the area of information exchange to share information with other bees by swinging dance. The more abundant nectar source, the greater probability to be selected, the more the onlooker bees. Then the onlooker bees search in the areas like the employed bee. The employed bee and onlooker bee choose the better nectar source position as the next generate solution according to the greedy rule.

Assume that the dimension of problem is D , the positions of nectar source corresponds to the points of solution space, the ith $(i = 1, 2, \cdots, NP)$ nectar source's quality is regarded as the fitness of solution fit_i , the number of soulution (NP, i.e. the nectar source's number), is total of the number of employed bees and onlooker bees. $\mathbf{X}_i = (x_{i1}, x_{i2}, \cdots x_{iD})$ represents the location of the ith nectar source, the location $\mathbf{X}_i$ is updated randomly in the d $d(d = 1, 2, \cdots, D)$ dimensions as follows.

$$x_{id} = L_d + rand(0,1) * (U_d - L_d) \tag{1}$$

where, x_{id} is the location of the ith bee in the dth dimension; L_d and U_d stand for the Lower and upper bounds of search space respectively.

At the beginning of the search ,the employed bee generates a new nectar source around the ith nectar source according to the formula 2.

$$v_{id} = x_{id} + rand(-1,1) * (x_{id} - x_{jd}) \tag{2}$$

where, d is a random integer which denotes that the employed bee searches in random dimension; $j \in \{1, 2, \cdots, NP\}$ ($j \neq i$) represents that it chooses a nectar source which is different from the ith nectar source in the NP nectar source; $rand(-1,1)$ is a random number with uniform distribution between 0 and 1.

When the fitness value of new nectar source $V_i = (v_{i1,} v_{i2}, \cdots, v_{id})$ is better than X_i, the V_i replaces X_i , otherwise, the ABC algorithm keeps X_i, and update the employed bees according to the formula (2). After updating, the employed bee feed information back to the onlooker bee. the onlooker bee chooses the employed bee to follow in roulette way according to the probability p_i , and determined the retention of nectar according to the same greedy method of the employed bee. p_i is computed as follows.

$$p_i = \frac{fit_i}{\sum_{j=1}^{SN} fit_j} \tag{3}$$

in which, fit_i stands for the fitness value of the ith nectar source.

In the processing of search, if X_i does not find the better the nectar source after reaching the threshold $limit$ and $trial$ times iteration, X_i would be given up and the employed bee would be changed into the scout. The scout generates randomly a new nectar source which replaces X_i in the search space. The above-mentioned process can be described as follows.

$$v_i^{t+1} = \begin{cases} L_d + rand(-1,1) * (U_d - L_d) & , \ trial \geq limit \\ v_i^t & , \ trial < limit \end{cases} \tag{4}$$

Generally, ABC algorithm takes the minimum optimization problem as an example, the fitness value of solution is estimated by formula (5).

$$fit_i = \begin{cases} 1/(1+f_i) & , \quad f_i \geq 0 \\ 1+abs(f_i) & , \ otherwise \end{cases} \tag{5}$$

where, the f_i is the function value of solution.

3 Artificial Bee Colony Using Opposition-Based Learning

According to the theory of probability, the each random generated candidate solution, compared with the opposite solution, has a probability of 50% away from the optimal solution of the problem. The literature [14] deduced the opposition-based learning strategy by math and proved the opposite candidate solution is close to the global optimal solution from the math theory. So the paper proposes a new method, called artificial bee colony using opposition-based learning, and to keep the advantage information of the employed bee an onlooker bee, we adopt a new update rule of the scout when the scout can not find the better nectar source after reaching the threshold $limit$ and $trial$ times iteration. It can increase the local exploration ability.

3.1 Opposite-Based Learning Strategy

Assuming in the kth iteration, the ith particle's optimal location and the opposite location are respectively p_i^m and op_i^m, and the values are respectively p_{id}^m and op_{id}^m in the mth dimension, then p_{id}^m is computed as

$$op_{id}^{m} = r_1 * a_d^m + r_2 * b_d^m - p_{id}^m \tag{6}$$

where r_1 and r_2 is a random number between 0 and 1, $p_{id}^m \in [a_d, b_d]$, $[a_d, b_d]$ is the dynamic boundary of the mth dimension space, $[a_d, b_d]$ is specified by

$$a_d^m = \min(x_{id}^m), b_d^m = \max(x_{id}^m) \tag{7}$$

The above formula (7) uses the dynamic boundary to replace the traditional fixed boundary, which is propitious to keep the previous search experience. In addition, if the opposite solution may jump out the boundary $[a_d, b_d]$, then reset according to the followings

$$\begin{cases} op_{id}^m = 2.0 * a_d(m) - op_{id}^m & if \quad op_{id}^m > x_{\max} \\ op_{id}^m = 2.0 * b_d(m) - op_{id}^m & if \quad op_{id}^m < x_{\min} \end{cases} \tag{8}$$

3.2 Update Rule of the Scout

To increase the local exploration ability of OBL-ABC, the scout is updated by formula (9).

$$x_{ik}^{t+1} = \begin{cases} rand(-1,1) * x_{jk}^t & , \quad trial \geq limit \\ x_{ik}^t & , \quad trial < limit \end{cases} \quad , \quad k = 1, 2, \cdots, D \tag{9}$$

where, x_{jk}^t is the location of the jth nectar source in kth dimension at the tth iteration, $j \in \{1, 2, \cdots, NP\}$ $j \neq i$ represents that it chooses a nectar source which is different from the ith nectar source in the NP nectar source; the new generated nectar is decided by the each dimension of the employed bee, some dimension of X_i may come from the some better location of the same nectar source.

3.3 The Steps of OBL-ABC

Step1: Initialization of parameters and nectar source, including NP, $limit$, CR, $trial = 0$;

Step2: Allocating the employed bee of NP nectar sources and updating by the formula (2). Then generating new nectar source V_i, and computing the fitness value of V_i according to formula (5); and then updating the employed bee by (6) and generating the opposite solution, also its fitness value is accurate by (5); finally, choosing the best fitness value according the greedy choose strategy, if V_i or opposite value is kept, set $trial = 0$,otherwise, $trial{+}{+}$;

Step3: according to roulette way, estimating the probability by formula (3) and updating according to the same greedy method of the employed bee.

Step4: Judging $trial \geq limit$, if it is satisfied, jump to Step 6, otherwise, go to Step5;

Step5: the scout generates the new nectar source according to formula (9);

Step6: if the algorithm satisfied the ending condition, output the optimum, otherwise , please turn to step 2 and continue.

4 Experimental Verifications

4.1 Test Functions

In order to verify the performance of OBL-ABC, We have chosen 6 well-known test functions as followings table 1, where functions f_1 to f_3 are unimodal functions, functions f_4 to f_6 are multimodal functions. The test functions are high dimensional multimodal functions, Most algorithms which are applied into these functions fall into local optimum.

Table 1. Benchmark functions

Function symbol	Function name	Search space	Optimum coordinate	optimum
f_1	Sphere	$[-100,100]$	$(0,\cdots,0)$	0
f_2	Schwefel.2.21	$[-100,100]$	$(0,\cdots,0)$	0
f_3	Quadric Noise	$[-1.28,1.28]$	$(0,\cdots,0)$	0
f_4	Rastrigin	$[-5.12,5.12]$	$(0,\cdots,0)$	0
f_5	Ackley	$[-32,32]$	$(0,\cdots,0)$	0
f_6	Griewank	$[-600,600]$	$(0,\cdots,0)$	0

4.2 Test Functions

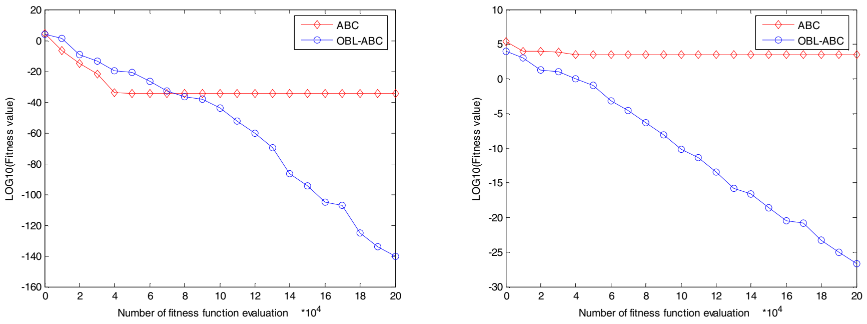
In the experiments ,the number of dimensions of all test functions in ABC and OBL-ABC is 30, evaluation number is 2.0×10^5 , the number of experiments is 100, the colony is 200, the number of employed bee and the onlooker bee is 100 respectively, when the individual search over 100 generation in the same nectar source, it will change to the scout. The comparison results are shown in the table 2. Fig.1 show the process comparison of function optimization.

Table 2. The comparison results with OBL-ABC and ABC on 6 benchmark functions

Function symbol	Algorithm	Mean	Std.Dev	Max Value	Min Value
f_1	OBL-ABC	1.05e-059	1.46e-058	5.27e-058	1.77e-073
	ABC	6.93e-016	6.93e-016	9.62e-016	4.42e-016
f_2	OBL-ABC	3.93e-010	3.51e-009	5.59e-009	1.82e-012
	ABC	3.69e+001	3.10e+001	4.98e+001	2.11e+001
f_3	OBL-ABC	1.60e-010	2.22e-009	8.03e-009	1.94e-038
	ABC	2.29e-001	2.94e-001	3.12e-001	8.67e-002
f_4	OBL-ABC	0.00	0.00	0.00	0.00
	ABC	7.42e-015	5.83e-014	4.26e-014	0.00
f_5	OBL-ABC	5.88e-016	0.00	5.88e-016	5.88e-016
	ABC	5.22e-014	3.65e-014	7.16e-014	3.96e-014
f_6	OBL-ABC	0.00	0.00	0.00	0.00
	ABC	1.91e-013	2.33e-012	4.22e-012	1.11e-016

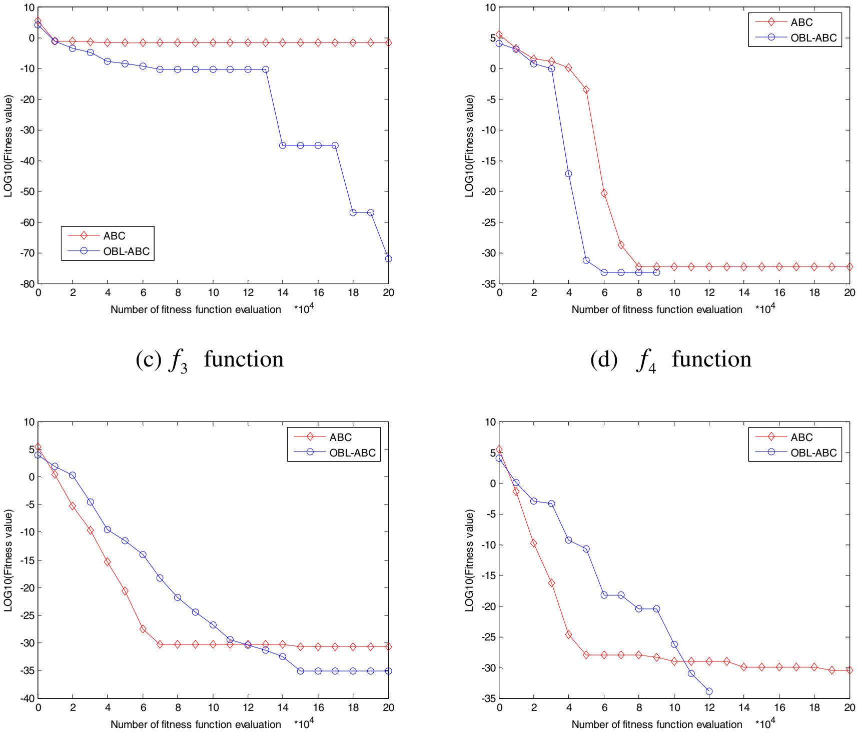
Table 2 presents results of ABC and OBL-ABC, the average fitness value , standard deviation, the maximum and minimum of OBL-ABC are further better than ABC's; OBL_ABC can find the optimal solution on f_4 and f_6 functions; especially, for Schwefel function whose optimal location is not in the 0 point and the solution is not 0, but the OBL-ABC can find the better solution for Schwefel function.

From Fig.1, OBL-ABC has stronger ability to search optimum and faster convergence velocity, and can reach the more perfect result after a few iteration times. However, the ABC falls into location optimum after a few iteration times.



(a) f_1 function (b) f_2 function

Fig. 1. The evolution trend of OBL-ABC and ABC



(c) f_3 function

(d) f_4 function

(e) f_5 function

(f) f_6 function

Fig. 1. (*continued*)

5 Conclusions

On the base of standard ABC, we introduce the opposite-based learning strategy, and present a new approach, called artificial bee colony using opposition-based learning. Our methods generates opposite solution for the employed bee and onlooker bee and keep the best location according greedy algorithm; in order to retain the advantage information of the employed bee and onlooker bee, a new update rule of the scout is proposed in the update processing of the scout. Finally, The simulation results demonstrated that OBL-ABC has promising convergence velocity and global optimum ability, compared with ABC on 6 famous test functions.

Acknowledgments. This work is supported by the National Natural Science Foundation of China under Grant(No. 61261039, and 61263029), Science Foundation of Jiangxi Province under Grant(No.20132BAB211031), Science and Technology Pillar Program of Jiangxi Province under Grant(No.20142BBG70034) and Technology plan projects of Nanchang city(No. 2013HZCG006, and 2013HZCG011).

References

1. Karaboga, D.: An idea based on honey bee swarn for numerical optimization. Erciyes University, Kayseri (2005)
2. Weifeng, G., Sanyang, L., Lingling, H.: Inspired Artificial Bee Colony Algorithm for Global Optimization Problems. Acta Electronica Sinica 40(2), 2396–2403 (2014)
3. Pei-Wei, C., Jeng-Shyang, P., Bin-Yih, L., Shu-Chuan, C.: Enhanced Artificial Bee Colony Optimization. International Journal of Innovative Computing, Information and Control 5(12), 5081–5092 (2009)
4. Amer, D., Bouaziz, A.: An artificial bee colony algorithm for image contrast enhancement. Swarm and Evolutionary Computation 16(6), 69–84 (2014)
5. Afandi, A.N., Miyauchi, H.: Improved artificial bee colony algorithm considering harvest season for computing economic dispatch on power system. IEEE Transactions on Electrical and Electronic Engineering 9(3), 251–257 (2014)
6. Ghani, A.A., Mohamad-Saleh, J.: Multiple-global-best guided artificial bee colony algorithm for induction motor parameter estimation. Turkish Journal of Electrical Engineering and Computer Sciences 22(3), 620–626 (2014)
7. Onder, B., Tasgetiren, M.F.: An artificial bee colony algorithm for the economic lot scheduling problem. International Journal of Production Research 42(4), 1150–1170 (2014)
8. Xiangtao, L., Minghao, Y.: Self-adaptive constrained artificial bee colony for constrained numerical optimization. Neural Computing and Applications 24(3), 723–734 (2014)
9. Tizhoosh, H.R.: Opposition-based Learning: a new scheme for machine intelligence. In: Proceedings of on Computational Intelligence for Modeling,Control and Automation, pp. 695–701 (2005)
10. Tizhoosh, H.R.: Reinforcement Learning Based on Actions and Opposite Actions. In: Proceedings of on Artificial Intelligence and Machine Learning, pp. 19–21 (2005)
11. Qing-zheng, X., Lei, W., Bao-min, H., et al.: Opposition-Based Differential Evolution Using the Current Optimum for Function Optimization. Journal of Applied Sciences 29(3), 308–315 (2011)
12. Haiping, M., Xieyong, R., Baogen, J.: Oppositional ant colony optimization algorithm and its application to fault monitoring. In: Proceedings of Chinese Control Conference, pp. 3895–3898 (2010)
13. Juan, L., Yi-wen, Z., Sen-lin, M.: Improved opposition-based shuffled frog leaping algorithm for function optimization problems. Application Research of Computers 30(3), 760–763 (2013)
14. Shahryar, R., Hamid, R.T., Magdy, M.A.S.: Opposition versus randomness in soft computing techniques. Applied Soft Computing 8(2), 906–918 (2008)

Theorem of Existence and Uniqueness of Fixed Points of Monotone Operators

Hui Luan[1] and Zhihong Xia[2]

[1] Faculty of Science, Nanchang Institute of Technology,
Nanchang China 330099
[2] Business Administration College, Nanchang Institute of Technology,
Nanchang China 330099

Abstract. Operator equation and the fixed point problem are an important component of nonlinear functional analysis theory. They are playing important role in solving nature and uniqueness problems about all kinds of differential equations and integral equations. Generally, the monotone operator has been defined with compactness, continuity and concavity and convexity in partially ordered Banach space. In this paper, without compactness and continuity, concavity and convexity of functions, a new fixed point theorem of increasing and decreasing operator and mixed monotone operator has obtained through introducing order-difference in the cone.

Keywords: cone, partial order, monotone operator, fixed point.

1 Introduction and Definition

The monotone operator has both important theoretical significance and wide applications. Generally, the monotone operator has been defined with compactness, continuity and concavity and convexity in partially ordered Banach space. In the literature [1,2], the compactness and continuity of monotone operator has been required; in the literature [3,4], the concavity and convexity of monotone operator has been required. In this paper, the definition of order-difference has been introduced in the cone; then the theorem of existence and uniqueness of fixed points of monotone operators have obtained without the above conditions.

In the following, let E be an arbitrary real Banach space, θ be zero element of E, non-empty convex closed set $P \subset E$ be cone, "$\leq$" be partial order from P, that is, $\forall x,\ \mathrm{y} \in \mathrm{E}$, and if $y - x \in P$, $x \leq y$. Suppose that $x,\ y \in E,\ x \leq y$, and definite that $[x,\ y] = \{z \in E \,|\, x \leq z \leq y\}$, the cone P can be considered to be normal. If exist $N > 0$, and $\theta \leq x \leq y \Rightarrow \|x\| \leq N \|y\|$, the minimum N can be defined as normal constant of P.

H. Sun et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 329, DOI: 10.1007/978-3-319-12286-1_2

Definition 1: Suppose that $D \subset E$, function $\varphi : [0,1) \times D \times D \to [0,1)$ can be considered to be meet the condition (Φ), If the following conditions are satisfied:

(i) $\varphi(0,x,y) = 0$, $\forall x,\ \mathrm{y} \in \mathrm{D} \times \mathrm{D}$;

(ii) $\forall (t,x,y) \in (0,1) \times D \times D$, $\exists\ W_0 \in D$,

then $0 \le \varphi(t,x,y) \le \varphi(t,w_0,w_0) < t$;

(iii) $\varphi^n(t,w_0,w_0) \to 0\ (n \to \infty)$, $\forall t \in [0,1)$.

and $\varphi^n(t,w_0,w_0) = \varphi^{n-1}(\varphi^n(t,w_0,w_0),w_0,w_0)$.

Definition 2: Let P be the cone of E, $\theta \le u \le v$, for $h \in P$, if exist $\exists M > 0$, $v \le Mh$, and $a = \inf\{\alpha | v \le \alpha h\}$, $b = \sup\{\beta | \beta h \le u\}$, $a-b$ could be called h-order-difference of u, v, and $d_h(u,v) = a-b$.

Note: In the definition 2, because $\{\beta | \beta h \le u\}$ includes element 0 at least and has a upper bound M, $\sup\{\beta | \beta h \le u\}$ may exist.

Lemma 1: P is a cone in E. $\theta \le u \le v$, $h \in P$, and $\exists M > 0, v \le Mh$. Then

(i) $d_h(u,v) \ge 0$, $d_v(u,v) \le 1$;

(ii) $d_h(u,v) = 0$, then $u = v$;

(iii) $d_h(u,v) \le d_h(u,w) + d_h(w,v)$, $\forall w \in [u,v]$. "=" is tenable if and only if $d_h(w,w) = 0$;

(iv) If $\theta \le u_1 \le u \le v$, then $d_h(u_1,v) \ge d_h(u,v)$;

(V) $d_h(ku,kv) = kd_h(u,v)$, $d_v(kv,v) = 1-k$, $\forall k \in [0,1]$.

Proof: Let $a = \inf\{\alpha | v \le \alpha h\}$, $b = \sup\{\beta | \beta h \le u\}$.

(i) In term of $\theta \le u \le v$ and $h \in P$, $\theta \le bh \le u \le v \le ah$ can be deduced, then $0 \le b \le a$. The definition 2 shows that $d_h(u,v) = a-b \ge 0$.

Let $a_1 = \inf\{\alpha | v \le \alpha v\}$, $b_1 = \sup\{\beta | \beta v \le u\}$.

Because $\theta \le u \le v$, $0 \le b_1 \le a_1 = 1$, then, $d_v(u,v) = a_1 - b_1 \le 1$.

(ii) Because $d_h(u,v) = 0$, $a = b$. Further more

$$\theta \le bh \le u \le v \le ah\text{, then } u = v.$$

(iii) $\forall w \in [u,v]$, $\theta \leq u \leq w \leq v \leq Mh$, therefore, $d_h(u,w)$ and $d_h(w,v)$ have meanings. Let $c = \inf\{\alpha | w \leq \alpha h\}$, $d = \sup\{\beta | \beta h \leq w\}$, in term of $w \leq ch$, $v \leq ah$ and $w \leq v$, $c \leq a$ can be known. (Actually, if $c > a$, the assumption of $w \leq v \leq ah$ and c is contradictory). Similarly, $b \leq d$ can be drawn. Then, $0 \leq b \leq d \leq c \leq a$, combined with Definition 2

$$d_h(u,v) = a-b \leq c-b+a-d = d_h(u,w)+d_h(w,v).$$

If $d_h(u,v)=d_h(u,w)+d_h(w,v)$, then $a-b=c-b+a-d$, that is, $c-d=0$, furthermore, $d_h(w,w)=0$ can be obtained. If $d_h(w,w)=0$, then $c-d=0$, furthermore,

$$d_h(u,v)=a-b=c-b+a-d = d_h(u,w)+ d_h(w,v).$$

(iv) Let $b_2 = \sup\{\beta | \beta v \leq u_1\}$, $b_2 \leq b$ can be drawn, therefore, $a-b_2 \geq a-b$, that is, $d_h(u_1,v) \geq d_h(u,v)$.

(V) From Definition 2, $d_h(ku,kv) = ka-kb = kd_h(u,v)$ and $d_v(kv,v) = 1-k$ can be obtained.

2 Main Conclusions

Theorem 1: P is the normal cone of E , $\theta \leq u_0 \leq v_0$. $A:[u_0,v_0] \to E$ is an increasing operator, and the following conditions are satisfied:

(i) $u_0 \leq Au_0$, $Av_0 \leq v_0$;

(ii) $\forall x,y \in [u_0,v_0]$, if $x \leq y$, then $d_{v_0}(Ax,Ay) \leq kd_{v_0}(x,y)$, constant $k \in (0,1)$. Therefore, the unique fixed point x^* of A may be existed in $[u_0,v_0]$, and for any initial value $x_0 \in [u_0,v_0]$, the iterative sequence $x_n = Ax_{n-1}\ (n=1,2,\cdots)$ must be converge to x^* .

Proof: Let $u_n = Au_{n-1}$, $v_n = Av_{n-1}$, $n=1,\ 2,\cdots$ (1)

From Condition (i), and A is an increasing operator, the following can be obtained:

$$u_0 \leq u_1 \leq u_2 \leq \cdots \leq u_n \leq \cdots \leq v_n \leq \cdots \leq v_2 \leq v_1 \leq v_0 \tag{2}$$

From the formula (1), (2) and Condition(ii), (i) of Lemma 1, the following can be obtained:

$d_{v_0}\left(u_n,v_n\right)=d_{v_0}\left(Au_{n-1},Av_{n-1}\right)\le kd_{v_0}\left(u_{n-1},v_{n-1}\right)\le\cdots\le k^n d_{v_0}\left(u_0,v_0\right)\le k^n$.

Therefore, while $n\to\infty$, $d_{v_0}\left(u_n,v_n\right)\to 0$. That is, for $\forall\varepsilon>0$, $K>0$ may exist, furthermore, while $n>K$, the following can be drawn:

$$d_{v_0}\left(u_n,v_n\right)<\varepsilon. \tag{3}$$

Letbe $a_n=\inf\left\{\alpha\left|v_n\le\alpha v_0\right.\right\}$, $b_n=\sup\left\{\beta\left|\beta v_0\le u_n\right.\right\}$. From the formula (3), the following can be obtained:

$$0\le a_n-b_n\le\varepsilon,\ b_n v_0\le u_n\le v_n\le a_n v_0.$$

And because P is normal, the following can be drawn while $n>K$:

$\left\|v_n-u_n\right\|\le N\left\|a_n v_0-b_n v_0\right\|\le N\varepsilon\left\|v_0\right\|$.(N is the normal constant of P)

The following can be obtained from the arbitrariness of ε:

$$v_n-u_n\to\theta\ \left(n\to\infty\right) \tag{4}$$

From the formula (2), the inequalities can be drawn: $\theta\le u_{n+p}-u_n\le v_n-u_n$, $\theta\le v_n-v_{n+p}\le v_n-u_n$. Furthermore, $\left\{u_n\right\}$ and $\left\{v_n\right\}$ are Cauchy consequences because of the formula (4) and the normality of P. Therefore, u^*, $v^*\in E$ and $u_n\to u^*$, $v_n\to v^*$, $u^*=v^*$. Let $x^*=u^*=v^*$, then $A\left(x^*\right)=x^*$.

$\forall x_0\in\left[u_0,v_0\right]$, the inequality $u_0\le x_0\le v_0$ may exist. The inequality $u_n\le x_n\le v_n$ ($x_n=Ax_{n-1}$, $n=1,2\cdots$) can be obtained with repeated substitution of A, therefore, $x_n\to x^*\left(n\to\infty\right)$. The above can give a conclusion: the unique fixed point x^* of A may be existed in $\left[u_0,v_0\right]$, and for any initial value $x_0\in\left[u_0,v_0\right]$, the iterative sequence $x_n=Ax_{n-1}\left(n=1,2,\cdots\right)$ must be converge to x^*.

Theorem 2: P is the normal cone of E , $\theta\le u_0\le v_0$. $A:\left[u_0,v_0\right]\to E$ is a decreasing operator, and the following conditions are satisfied:

(i) $u_0\le Au_0$, $Av_0\le v_0$;

(ii) $\forall x,y\in\left[u_0,v_0\right]$, if $x\le y$, then $d_{v_0}\left(Ay,Ax\right)\le kd_{v_0}\left(x,y\right)$, constant $k\in\left(0,1\right)$. Therefore, the unique fixed point x^* of A may be existed in $\left[u_0,v_0\right]$, and for any initial value $x_0\in\left[u_0,v_0\right]$, the iterative sequence $x_n=Ax_{n-1}\left(n=1,2,\cdots\right)$ must be converge to x^*.

Proof: Let $B = A^2$, the following can be proved:

$B:[u_0, v_0] \to E$ is an increasing operator, and

$$u_0 \le Bu_0,\ \mathrm{B}v_0 \le v_0. \tag{5}$$

$\forall x, y \in [u_0, v_0]$, if $x \le y$, from Condition (ii) the following can be obtained:

$$d_{v_0}(Bx, By) = d_{v_0}\left(A^2x, A^2y\right) \le kd_{v_0}(Ay, Ax) \le k^2 d_{v_0}(x, y) \tag{6}$$

The conditions of Theorem 1 has satisfied from the formulas (5) and (6), therefore, the conclusions of Theorem 2 can be established.

Theorem 3 and Theorem 4 are introduced in the following, and their proofs are similar, so we can only prove Theorem 4.

Theorem 3: P is the normal cone of E, $\theta \le u_0 \le v_0$. $\varphi:[0,1)\times[u_0,v_0]\times[u_0,v_0] \to [0,1)$ can be satisfied with Condition (Φ). $A:[u_0,v_0] \to E$ is an increasing operator, and the following conditions are satisfied:

(i) $u_0 \le Au_0,\ Av_0 \le v_0$;

(ii) $\forall x, y \in [u_0, v_0]$, if $x \le y$, then

$td_{Ay}(Ax, Ay) \le \varphi(t, x, y)d_y(x, y), \forall\ t \in (0,1)$. Therefore, the unique fixed point x^* of A may be existed in $[u_0, v_0]$, and for any initial value $x_0 \in [u_0, v_0]$, the iterative sequence $x_n = Ax_{n-1}\ (n = 1, 2, \cdots)$ must be converge to x^*.

Theorem 4: P is the normal cone of E, $\theta \le u_0 \le v_0$. $\varphi:[0,1)\times[u_0,v_0]\times[u_0,v_0] \to [0,1)$ can be satisfied with Condition (Φ). $A:[u_0,v_0] \to E$ is a mixed monotone operator, and the following conditions are satisfied:

(i) $u_0 \le A(u_0, v_0),\ A(v_0, u_0) \le v_0$;

(ii) $\forall x, y \in [u_0, v_0]$, if $x \le y$, then

$td_{A(y,x)}(A(x, y), A(y, x)) \le \varphi(t, x, y)d_y(x, y),\ \forall\ t \in (0,1)$. Therefore, the unique fixed point x^* of A may be existed in $[u_0, v_0]$, and for any initial value $x_0 \in [u_0, v_0]$, the iterative sequence $x_n = A(x_{n-1}, y_{n-1})$, $y_n = A(y_{n-1}, x_{n-1}),\ (n = 1, 2, \cdots)$ must be converge to x^*.

Proof: Let $u_n = A(u_{n-1}, v_{n-1})$, $v_n = A(v_{n-1}, u_{n-1})$, $n = 1, 2, \cdots$, the following can be obtained:

$$u_0 \le u_1 \le u_2 \le \cdots \le u_n \le \cdots \le v_n \le \cdots \le v_2 \le v_1 \le v_0. \tag{7}$$

From Condition (ⅱ) and (Φ), $t \in (0,1)$, the following can be drawn:

$$\begin{aligned} td_{v_n}(u_n, v_n) &\le \varphi(t, u_n, v_n) d_{v_{n-1}}(u_{n-1}, v_{n-1}) \le \varphi(t, w_0, w_0) d_{v_{n-1}}(u_{n-1}, v_{n-1}) \\ &\le \cdots \le \varphi^n(t, w_0, w_0) d_{v_0}(u_0, v_0) \le \varphi^n(t, w_0, w_0) \end{aligned} \tag{8}$$

Letbe $a_n = \inf\{\alpha | v_n \le \alpha v_n\}$, $b_n = \sup\{\beta | \beta v_n \le u_n\}$. Therefore, while $n \to \infty$, $d_{v_n}(u_n, v_n) \to 0$, and $a_n - b_n \to 0 (n \to \infty)$,

$$\theta \le bv_n \le u_n \le v_n \le av_n. \tag{9}$$

The following can be obtained from the normality of P and the formula (9): $\|v_n - u_n\| \le N\|av_n - bv_n\| = N(a-b)\|v_n\|$, ($N$ is the normal constant of P)

And $[u_0, v_0]$ is normal-bounded, therefore, $v_n - u_n \to \theta \ (n \to \infty)$.

The inequalities of $\theta \le u_{n+p} - u_n \le v_n - u_n$ and $\theta \le v_n - v_{n+p} \le v_n - u_n$ can be obtained from the formula (7). Furthermore, $\{u_n\}$ and $\{v_n\}$ are Cauchy consequences. Therefore, u^*, $v^* \in E$ and $u_n \to u^*$, $v_n \to v^*$, $u^* = v^*$. Let $x^* = u^* = v^*$.

Because the proof of uniqueness is similar to Theorem 1, it is omitted.

The following theorem can be obtained from the proof of Definition 2, Lemma 1 and Theorem 4.

Theorem 5: P is the normal cone of E, $\theta \le u_0 \le v_0$. $u_0 \le u_1 \le u_2 \le \cdots \le u_n \le \cdots \le v_n \le \cdots \le v_2 \le v_1 \le v_0$. If one of the following conditions is satisfied:

(ⅰ) $d_{v_n}(u_n, v_n) \le kd_{v_{n-1}}(u_{n-1}, v_{n-1})$, $n = 1, 2, \cdots$, constant $k \in (0,1)$;

(ⅱ) $d_{v_n}(u_n, v_n) \le \alpha_n$, $n = 1, 2, \cdots$, the series $\{\alpha_n\}$ is convergent, and $\alpha_n \to 0 \ (n \to \infty)$. Therefore, $u_n \to u^*$, $v_n \to v^*$, and $u^* = v^*$ while $n \to \infty$.

References

1. Guo, D.J., Lakshmikantham, V.: Coupled fixed points of npnlinear operators with applications. Nonlinear Analysis, TMA 11(5), 623–632 (1978)
2. Sun, Y.: A fixed-point theorem for mixed monotone operators with applications. J. Math. Anal. Appl. 156, 240–252 (1991)
3. Zhang, M.Y.: Fixed point theorems of convex concave mixed monotone operators and applications. J. Math. Anal. Appl. 339, 970–981 (2008)
4. Wu, Y.S., Li, G.Z.: On the fixed point existence and uniqueness theorems of mixed monotone operators and applications. J. A. Math. S., Chinese Series 46(1), 161–166 (2003)
5. Xu, S., Zeng, C., Zhu, C.: Existence and Uniqueness for the Fixed Points of φ Concave-(-ψ) Convex Mixed Monotone Operators and Its Applications. J. Acta Mathematica Sinica 48(6), 1055–1064 (2005)
6. Dajun, G.: Method of partial ordering in nonlinear analysis. Shandong Science & Technology Press (2000)
7. Zhu, C.X.: Several nonlinear operator problems in the Menger PN space. J. Nonlinear Analysis 65, 1281–1284 (2000)
8. Hong, S.H.: Fixed points for mixed monotone multivalued operators in Banach spaces with applications. J. Math. Anal. Appl. 337, 333–342 (2008)

Adaptive Sampling Detection Based Immune Optimization Approach and Its Application to Chance Constrained Programming

Kai Yang[1] and Zhuhong Zhang[2]

[1] College of Computer Science, Guizhou University (550025), Guiyang, China
csc.kyang@gzu.edu.cn
[2] College of Electronics and Information, Guizhou University (550025), Guiyang, China
sci.zhzhang@gzu.edu.cn

Abstract. This work investigates a bio-inspired adaptive sampling immune optimization algorithm to solve linear or nonlinear chance-constrained optimization problems without any noisy information. In this optimizer, an efficient adaptive sampling detection scheme is developed to detect individual's feasibility, while those high-quality individuals in the current population can be decided based on the reported sample-allocation scheme; a clonal selection-based time-varying evolving mechanism is established to ensure the evolving population strong population diversity and noisy suppression as well as rapidly moving toward the desired region. The comparative experiments show that the proposed algorithm can effectively solve multi-modal chance-constrained programming problems with high efficiency.

Keywords: Chance-constrained programming, immune optimization, feasibility detection, sample-allocation, multimodality.

1 Introduction

Chance constrained programming (CCP), originally introduced by Charnes and Cooper[1], is a kind of stochastic programming, including at least one chance constraint or probabilistic inequality. Solving such type of problem involves two crucial factors: (i) how to design efficient computational models for chance constraints, and (ii) probing into effective optimizers for rapidly finding the optimal solution. Even if CCP has been intensively studied by mathematical researchers[1-3], few optimization approaches are reported for real-world engineering chance-constrained optimization problems such as control system design and electric power system dispatching. The main difficulty includes: (i) identifying individual's feasibility becomes difficult when stochastic variables are without any distribution information, and (ii) the feasible region is generally non-convex. Usually, stochastic simulation and model approximation are alternative for constraint handling, e.g., sample average approximation[2] and scenario approximation[3]. Monte Carlo simulation is a competitive sample approximation approach, but needs the same large sample size for each individual, i.e., static

H. Sun et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 329, DOI: 10.1007/978-3-319-12286-1_3

sampling. Especially, adaptive sampling[4] becomes increasingly popular in the context of stochastic optimization, as different individuals are attached different sample sizes. Such type of approach can reduce heavily computational cost and help for rapidly finding the optimal solution. From the viewpoint of optimization, although conventional numerical methods behave worse for nonlinear CCP, some advanced stochastic approaches have presented their potentials in the branch of intelligent optimization[1,4-5]. For example, Liu[1] investigated how artificial neural networks approached chance constraints, and used genetic algorithms to search the desired solution; Poojari et al.[5] developed two similar genetic algorithms (SSGA-A and SSGA-B) with the same static sampling scheme. These achievements are valuable for CCP, but need to make some improvements on computational complexity.

Immune optimization as a hot topic in the field of artificial immune systems is gaining great attention among intelligent researchers. Whereas many immune optimizers presented their superiority in solving complex static or dynamic multimodal engineering problems[6], they are rarely applied to CCP problems. Recently, Zhao et al.[7] developed a hybrid immune optimization approach with the static sampling strategy to solve nonlinear CCP problems, in which the neural network was used to approximate the stochastic functions, and meanwhile two operators of double cloning and double mutation were designed to accelerate the process of immune evolution. However, such approach needs a lot of runtime to simulate specific stochastic functions. Our recent work[4,8] also studied two adaptive sampling immune optimization algorithms and their theory for nonlinear joint and non-joint CCP. In such work, one of concerns is to investigate how to guarantee that all empirical individuals with different importance gain different sample sizes. However, some improvements need to be done, e.g., sampling efficiency and individual's feasibility detection. Based on such consideration, we in the current work proposed an adaptive sampling detection-based immune optimization approach (ASDIOA) to find CCP's optimal solution, especially an efficient adaptive sampling detection approach was studied based on Hoeffding's inequality[2]. Compared to our previous optimizers for CCP, ASDIOA is more efficient and can achieve effective solution search.

2 Problem Statement and Preliminaries

Consider the following nonlinear chance-constrained programming problem:

$$\text{(CCP)} \qquad \min E[f(\boldsymbol{x},\xi)]$$
$$s.t.\begin{cases} \Pr\{G_i(\boldsymbol{x},\xi)\le 0\}\ge \alpha_i, & 1\le i\le I \\ g_j(\boldsymbol{x})\le 0,\ 1\le j\le J, & h_k(\boldsymbol{x})=0,\ 1\le k\le K, \end{cases}$$

with bounded and closed domain D in R^p, design vector $\boldsymbol{x}$ in D, random vector ξ in R^q and confidence levels α_i in (0,1), where E [.] and Pr{.} are the operators of expectation and probability respectively; $f(\boldsymbol{x}, \xi)$ and G_i $(\boldsymbol{x}, \xi)$ are the stochastic objective and constraint functions respectively; $g_j(\boldsymbol{x}, \xi)$ and $h_k(\boldsymbol{x})$ are the deterministic constraint functions. If a candidate satisfies all the above constraints, it is called a feasible solution, and an infeasible solution otherwise. Introduce the constraint violation function to check whether candidate $\boldsymbol{x}$ is feasible:

$$\begin{aligned}\Gamma(\boldsymbol{x}) = I^{-1}\sum_{i=1}^{I}\max\{\alpha_i - \Pr\{G_i(\boldsymbol{x},\xi) > 0\},0\} \\ + J^{-1}\sum_{j=1}^{J}\max\{g_j(\boldsymbol{x}),0\} + K^{-1}\sum_{k=1}^{K}|h_l(\boldsymbol{x})|.\end{aligned} \tag{1}$$

Obviously, $\boldsymbol{x}$ is feasible only when $\Gamma(\boldsymbol{x}) = 0$.

Chen[9] developed a sample-allocation method (OCBA) to allocate a total sampling size of population to different individuals so as to find the best solution. In the present work, OCBA is adopted to decide sampling sizes of m empirically feasible candidates with a total sample size of T. More precisely, let A represent a population of the above m candidates with sample size T, and N_j^l the sample size of the j-th candidate at the moment l. σ_i^l denotes the variance of observations for candidate i with sample size N_i^l, and c stands for the best candidate whose empirically objective value is smallest in A. $\delta_{c,i}$ is the Euclidian distance between candidates c and i in the design space. Thereafter, OCBA can be reformulated below:

Step 1. Set $l \leftarrow 0$. Each candidate in A creates m_0 observations with $N_1^l = N_2^l = \ldots = N_m^l = m_0$;

Step 2. Decide the best candidate c through the empirical means of all candidates in A, and calculate $\sigma_i, \delta_{c,i}, 1 \le i \le m$;

Step 3. If $\sum_{i=1}^{m} N_i^l > T$, this procedure ends, and outputs the empirical means of candidates in A; otherwise, go to Step 4;

Step 4. Increase the computing budget by Δ, and decide $N_i^{l+1}, i = 1,2,\ldots, m$ through the following formula

$$\frac{N_i}{N_j} = \left(\frac{\sigma_i/\delta_{c,i}}{\sigma_j/\delta_{c,j}}\right)^2, i,j \in \{1,2,\cdots,m\}, i \neq j \neq m, N_c = \sigma_c\sqrt{\sum_{i=1,i\neq c}^{m}\frac{N_i^2}{\sigma_i^2}}\ ; \tag{2}$$

Step 5. If $N_i^{l+1} > N_i^l$, then $N_i^{l+1} \leftarrow N_i^l + \max\{0, N_i^{l+1} - N_i^l\}$, and go to Step 2.

Theorem 1 (Hoeffding's Inequality)[2]. If $X_1, X_2, \ldots, X_n$ are iid random variables with $a \le X_i \le b$ and mean μ, then

$$|\overline{X}_n - \mu| \le (b-a)\sqrt{\frac{1}{2n}\ln\left(\frac{2}{\delta}\right)},$$

with probability at least 1-δ.

3 Adaptive Sampling Detection for Chance Constraints

Sample average approximation is a usual approach handling chance-constrained programming[2], in which Monte Carlo simulation is used to estimate the probability of a chance constraint under a given sample size. However, when intelligent optimization approaches solve an approximate model of such kind of problem, it is impossible to avoid identifying whether a candidate is feasible or not. If the candidate is empirically feasible, it is desired to attach a large sample size, as its empirical objective mean is required to approach the theoretical mean as possible; conversely, it is not necessary to provide an empirically infeasible candidate with a large sample size. To this point, we develop a Hoeffding's inequality-based sampling detection scheme to decide the sample size for a candidate, and meanwhile detect whether such candidate satisfies a given chance constraint. More precisely, let T be a maximal sampling size of candidate $\boldsymbol{x}$; α denotes the confidence level of a given chance constraint, satisfying $\Pr\{g(\boldsymbol{x},\xi)\leq 0\}\geq\alpha$; $p_n(\boldsymbol{x})$ represents the probability estimate at the moment n through Monte Carlo simulation. Thereafter, the sampling scheme, simply say adaptive sampling detection, is formulated as follows:

Step 1. Input parameters: initial sample size m_0, sampling amplitude λ, maximal sampling size T;
Step 2. Set $m=m_0$, $s=m_0$; calculate the probability estimate $p_n(\boldsymbol{x})$ with m_0 observations of candidate $\boldsymbol{x}$;
Step 3. End this procedure only when $m>T$ or $|p_n(x)-\alpha|>\sqrt{ln(2/\delta)/(2m)}$;
Step 4. Set $s\leftarrow\lambda s$;
Step 5. Create s observations of the candidate, and decide the successful rate, $r_n\leftarrow w/s$, where w denotes the number that the inequality, $g(\boldsymbol{x},\xi)\leq 0$, is true for the observations;
Step 6. Update the probability estimate, i.e., $p_n(\boldsymbol{x})\leftarrow(p_n(\boldsymbol{x})\times m+r_n s)/(m+s)$;
Step 7. Set $m\leftarrow m+s$, and return to Step 3.

When all the candidates in a given population are attached the same maximal sample size T, the above procedure follows that different candidates get different sample sizes. Worse candidates gain smaller sample sizes. Here, candidate $\boldsymbol{x}$ is called empirically feasible if $p_n(\boldsymbol{x})\geq\alpha$ and the above deterministic constraints are satisfied, and infeasible otherwise.

4 Algorithm Formulation and Design

The clonal selection theory explains a biological phenomenon which antibodies respond to an antigen. It also hints a learning mechanism that the antigen can be gradually eliminated by some antibodies. To utilize such theory to design ASDIOA for CCP, a real-coded candidate is regarded as an antibody; the problem itself is viewed as the antigen. Based on the above OCBA, adaptive sampling detection and bio-immune inspirations, ASDIOA can be described in detail below:

Step 1. Input parameters: population size N, initial sample size m_0, sampling amplitude λ, computing budget Δ and maximal clonal size $C_{\max}$;

Step 2. Set $n \leftarrow 1$. Generate an initial population A_n of N random antibodies;

Step 3. Calculate the probability estimate of each chance constraint in the above CCP for each antibody in A_n through the adaptive sampling detection with a maximal sample size, $T=m_0[(n+1)^{1/2}+1]$; detect whether each antibody in A_n is empirically feasible through equation (1);

Step 4. Divide A_n into empirically feasible sub-population B_n and infeasible sub-population C_n;

Step 5. Allocate the sample size of population, $T=m_0|B_n|log(n+2)$, to each antibody in B_n through the above OCBA, and calculate the empirical objective values of all the antibodies;

Step 6. Each antibody in B_n and C_n proliferates $cl(\boldsymbol{x})$ clones with $cl(\boldsymbol{x})=$ *round* $(C_{\max}/(\Gamma(\boldsymbol{x})+1)+1)$, which creates a clonal population D_n, where $round(z)$ is a maximal integer not beyond z;

Step 7. Each clone in D_n shifts its genes through the conventional Gaussian mutation with a mutation rate $p_m=1/(\Gamma_{\max}-\Gamma(\boldsymbol{x})+1)$, where $\Gamma_{\max}$ denotes the maximal of constraint violations for all the clones in D_n; thereafter, all mutated clones constitute E_n and execute evaluation through Steps 3 to 5 with $n=1$;

Step 8. Execute comparison between antibodies and their clones in $A_\mathrm{n} \cup E_\mathrm{n}$. If some antibody in A_n is inferior to the best of its clones, such antibody is replaced by the best clone. This creates a new population A_{n+1};

Step 9. If the termination criterion is not satisfied, then set $n \leftarrow n+1$ and go to Step 3; otherwise, output the best antibody viewed as the optimal solution.

In the above algorithm, after checking feasibility for each antibody in A_n in step 3, two sub-populations evolve respectively toward different directions through steps 6 to 8. The empirical objective values of antibodies in step 5 are decided through dynamically allocating a time-dependent population sample size to empirically feasible antibodies. Steps 6 and 7 urge high-quality antibodies to produce multiple clones with small mutation rates. Obviously, those survival and better antibodies can gain larger sample sizes when gradually increasing the iteration number n. Therefore, ASDIOA is a dynamically sampling optimizer.

Additionally, ASDIOA's computational complexity can be decided by Steps 3 to 5 and 7 to 8. Step 3 needs at most $I \times T$ times to create samples in the worst case, and thereby the computational complexity is $O(IN(m_0\ (n+1)^{1/2}+m_0))$. In addition, Step 4 divides A_n into two sub-populations with N executions. Step 5 calculates the empirical objective means of the empirically feasible antibodies in B_n with a total of evaluations, $m_0|B_n|log(n+2)$, and hence the complexity is $O(Nm_0 log(n+2))$ in the worst case. Step7 executes mutation with at most $N(1+C_{\max})$ times. In Step 8, each clonal sub-population needs to execute comparison with at most $C_{\max} log C_{\max}$ times, and thus the complexity is at most $O(N\mathrm{C}_{\max}log\mathrm{C}_{\max})$. Therefore, ASDIOA's computational complexity can be given by

$$\begin{aligned} O_c &= O(IN(m_0\sqrt{n+1}+m_0))+O(Nm_0\log(n+2))+O(NC_{\max}\log C_{\max}) \\ &= O(N(Im_0\sqrt{n+1}+C_{\max}\log C_{\max})). \end{aligned} \tag{3}$$

5 Numerical Experiments

In this experimental study, four representative intelligent algorithms suitable for CCP, i.e., one neural network-based particle swarm optimization (HPSO)[1], two competitive steady genetic algorithms (SSGA-A and SSGA-B)[5] and one recent noisy immune optimization approach (NIOA)[4], are picked up to compare to ASDIOA by means of the following test examples. Our experiments are executed on a personal computer with CPU/3GHz and RMB/2 GB and also VC++. Especially, the three approaches of HPSO, SSGA-A and SSGA-B are static sampling optimization approaches with the same fixed sample size for each individual, whereas NIOA is an adaptive sampling optimization approach with a dynamic sample size for each individual. Their parameter settings are the same as those in their corresponding literature except their evolving population sizes. All the above algorithms take their population sizes 40, while respectively executing 30 times on each test problem. Their same termination criterion is that the total of evaluations of individuals during evolution is 10^7. Especially, HPSO is a BP neural network-based optimization approach, in which the total of training sample size is set as 10^7. In ASDIOA, after experimental tuning we take m_0=30, λ=1.5, Δ=20 and $\mathrm{C}_{\max}$=2. In order to effectively execute comparison between the algorithms, each of those solutions, gotten by them is required to re-evaluate with the sample size 10^6. Here, we give a test criterion to examine whether the solutions satisfy the chance constraints; in other words, let Λ denote an empirically infeasible solution set with size M, and $p^i(\boldsymbol{x}_l)$ represents the probability estimate of the *i-th* chance constraint as in Section 2 for $\boldsymbol{x}_l$ in Λ. The test criterion is designed as follows:

$$IAE = M^{-1}\sum_{l=1}^{M}\sum_{i=1}^{I}\left|p^i(\boldsymbol{x}_l)-\alpha_i\right| \tag{4}$$

Example 1. Feed Mixer Problem[5]

$$\min E\,[24.55x_1+26.75x_2+39.0x_3+40.50x_4+\xi]$$

$$s.t\begin{cases} x_1+x_2+x_3+x_4=1,\ \ 2.3x_1+5.6x_2+11.1x_3+1.3x_4\ge 5, \\ \Pr\{\eta_1x_x+\eta_2x_2+\eta_3x_3+\eta_4x_4\ge 21\}\ge 0.8 \\ x_1,x_2,x_3,x_4\ge 0,\ \ \eta_1\sim N(12,0.2809^2),\eta_2\sim N(11.9,0.1936^2), \\ \eta_3\sim N(41.8,20.25^2),\eta_4\sim N(52.1,0.6241^2),\ \ \xi\sim N(0,1). \end{cases}$$

Although this is a linear CCP with 4 decision variables, the five random variables influence the process of solution search seriously. It can be transformed into a deterministic optimization problem with the theoretical minimum 30.30 at the point (0.002, 0.733, 0.056, 0.209). However, in order to examine the performances of the above algorithms, we directly solve such problem. After respectively running 30 times, the algorithms can get their solution sets which create statistical results listed in Table 1, while Figure 1 below shows the box plot of the statistical results and Figure 2 displays their average search curves.

Table 1. Comparison of statistical result for problem 1

Algorithm	*Max*	*Min*	*Mean*	*Std.Dev*	*CI*	*IAE*	*FR*	*AR*(s)
HPSO	35.85	23.23	32.87	2.99	[31.80,33.94]	0.44	0%	15.5
SSGA-A	30.71	30.24	30.45	0.13	[30.41,30.50]	0.05	0%	7.2
SSGA-B	30.75	30.14	30.41	0.13	[30.36,30.45]	0.05	3%	7.3
NIOA	30.36	30.02	30.24	0.08	[30.21,30.27]	0.07	7%	7.3
ASDIOA	**30.51**	**30.17**	**30.33**	**0.09**	**[30.30,30.36]**	**4.18×10^{-4}**	**97%**	**6.7**

CI represents the confidence interval of empirically objective means for the 30 solutions acquired; *IAE* is computed through equation (4) for empirically infeasible solutions gotten by a given algorithm; *FR* stands for the rate of empirically feasible solutions among all the gotten solutions; *AR* is the average runtime after 30 runs for a given algorithm.

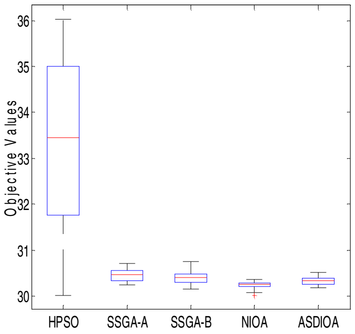


Fig. 1. Box-plot of problem 1

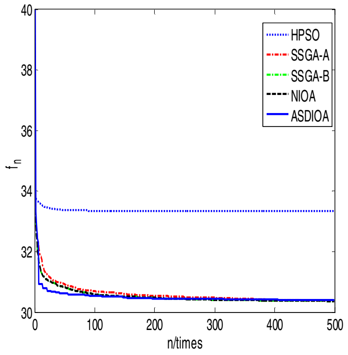


Fig 2. Average search curves of problem1

In Table 1, the values of *FR* listed in the eighth column hints that HPSO and SSGA-A can not find feasible solutions and that SSGA-B and NIOA can only get a few feasible solutions, whereas the solutions gotten by ASDIOA are almost feasible. On the other hand, the values of *IAE* in the seventh column show that ASDIOA only causes the smallest constraint violation for the chance constraints, which indicates that the adaptive sampling detection as in section 3 can effectively handle such constraints; those compared approaches are difficult in solving such kind of constraint. It seems to be true that all the algorithms but HPSO have almost the same solution quality through the values as in the columns 4 and 6. In fact, as associated to their constraint violations (*IAE*), the sizes of their solutions and the theoretical minimum, ASDIOA has the best solution quality obviously; especially, the values of *CI* hint that the minimum, 30.30, is included in the narrow confidence interval obtained by ASDIOA, but other algorithms are difficult. Apart from HPSO, the other three algorithms can gain the better solution qualities.

Through the columns 2, 3, 5 and 6, we can get the conclusion that all the algorithms but HPSO have relatively stable search performances. The statistical box-plots of the empirical objective values in Figure 1, acquired by the algorithms after 30 executions illustrates a fact that NIOA and ASDIOA can obtain similar effects superior to those gained by other algorithms, and meanwhile their objective values cover small scopes. By Figure 2, we also note that ASDIOA is convergent, and HPSO can only achieve local solution search; relatively, ASDIOA is a rapid search procedure. Lastly, the values of *AV*, listed in the ninth column present clearly that all the algorithms but HPSO have high search efficiency; HPSO spends the more runtime to solve the above problem than each of other algorithms.

Example 2. Multi-modal Problem

$$\max\ E\,[\sum_{n=1}^{3} x_i \cdot \sin(i\pi x_i) + \xi]$$

$$s.t.\begin{cases}\Pr\{\eta_1 x_1 + \eta_2 x_2 + \eta_3 x_3 \le 10\} \ge 0.7,\\ \Pr\{\eta_4 x_1 + \eta_5 x_2 + \eta_6 x_3 \le 100\} \ge 0.8,\\ x_1, x_2, x_3 \ge 0\ ,\ \xi \sim N(0,1), \eta_1 \sim U(0.8,1.2),\quad \eta_2 \sim U(1,1.3),\\ \eta_3 \sim U(0.8,1)\ ,\ \eta_4 \sim N(1,0.5)\ ,\ \eta_5 \sim E\text{xp}(1.2)\ ,\ \eta_6 \sim L\text{og}(0.8,0.6).\end{cases}$$

This is a multimodal chance-constrained programming problem gotten through modifying the static multimodal optimization problem [10], where three decision variables and four random variables are included. The main difficulty of solving such problem involves two aspects: (i) the original static problem has multiple local optima, and (ii) the above problem involves in multiple kinds of random variables. Like the above experiment, the approaches can get their statistical results (Table 2) and performance curves (Figures 3 and 4) after 30 runs.

Table 2. Comparison of statistical result for problem 2

Algorithm	Max	Min	Mean	Std.Dev	CI	IAE	FR	AR(s)
HPSO	11.09	-4.41	3.11	4.20	[1.60,4.61]	0.06	0.80	10.0
SSGA-A	10.08	7.51	9.07	0.59	[8.86,9.28]	0.03	0.80	7.1
SSGA-B	10.10	6.47	8.85	0.74	[8.58,9.12]	0.03	0.87	7.1
NIOA	9.88	9.23	9.55	0.15	[9.50,9.61]	0.06	0.67	7.1
ASDIOA	**10.08**	**9.47**	**9.82**	**0.18**	**[9.75,9.86]**	**2.17×10^{-3}**	**0.93**	**6.3**

Following Table 2, the values of *FR* show that the above algorithms can all find some feasible solutions after 30 runs; relatively, ASDIOA is more effective. As related to the data listed in the seventh column, we notice that the approaches can almost handle the above chance constraints; especially, ASDIOA's adaptive sampling detection has presented its prominent performance with the aspect of dealing with chance constraints. On the other hand, we can acquire significantly different solution qualities for the algorithms by means of the statistical results given in columns 2 to 6. In other words, ASDIOA's solution quality is significantly superior to those acquired by other approaches; NIOA is secondary, and HPSO is worst (see Figure 3). We also observe that ASDIOA has presented its strong and stable evolving ability of searching the optima (see Figure 4), as it can get the largest objective value and the narrowest confidence interval.

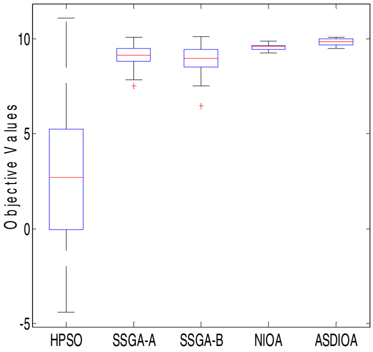


Fig 3. Box-plot of problem 2

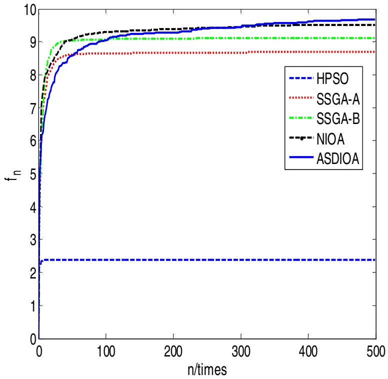


Fig 4. Average search curves of problem 2

Figure 4 indicates that NIOA and ASDIOA are convergent but other algorithms get into local search. This illustrates that the adaptive sampling schemes presented in NIOA and ASDIOA can help these two algorithms improve their solution qualities. Additionally, we can get the same conclusion on performance efficiency as that given in Example 1, namely ASDIOA spends the least time to execute the process of solution search but HPSO is worst when doing so.

6 Conclusions

With the increasing requirement of handling uncertain optimization problems, CCP will become increasingly popular in the field of intelligent optimization. Thus, inspired by the dynamic characteristics and mechanisms of the immune system, this work focuses on probing into a bio-inspired adaptive immune optimization algorithm for chance-constrained programming problems with any distributions. Especially, an efficient adaptive sampling detection scheme is developed to handle chance constraints, while the existing OCBA is used to make high-quality individuals gain large sampling sizes. Such algorithm is an optimizer capable of effectively executing noisy suppression, adaptive sample-allocation and chance constraint handling. The experimental results hint that the proposed approach is a competitive, effective and efficient optimizer.

Acknowledgement. This research is supported by Doctoral Fund of Ministry of Education of China (20125201110003), National Natural Science Foundation (61065010) and EU FP7 HAZCEPT (318907).

References

1. Liu, B.: Theory and practice of uncertain programming. STUDFUZZ, vol. 239. Springer, Heidelberg (2009)
2. Luedtke, J., Ahmed, S.: A sample approximation approach for optimization with probabilistic constraints. SIAM J. Optim. 19, 674–699 (2008)
3. Nemirovski, A., Shapiro, A.: Scenario approximations of chance constraints. In: Calafior, G., Dabbene, F. (eds.) Probabilistic and Randomized Methods for Design under Uncertainty, pp. 3–48. Springer, London (2005)
4. Zhang, Z.H.: Noisy immune optimization for chance-constrained programming problems. Applied Mechanics and Materials 48, 740–744 (2011)
5. Poojari, C.A., Varghese, B.: Genetic algorithm based technique for solving chance constrained problems. European Journal of Operational Research 185(3), 1128–1154 (2008)
6. Dasgupta, D., Yu, S., Nino, F.: Recent advances in artificial immune systems: models and applications. Applied Soft Computing 11(2), 1574–1587 (2011)
7. Zhao, Q., Yang, R., Duan, F.: An immune clonal hybrid algorithm for solving stochastic chance-constrained programming. *Journal of Computational Information Systems 8(20), 8295–8302 (2012)*
8. Zhang, Z.H., Wang, L., Liao, M.: Adaptive sampling immune algorithm solving joint chance-constrained programming. Journal of Control Theory and Application 11(2), 237–246 (2013)
9. Chen, C.H.: Efficient sampling for simulation-based optimization under uncertainty. In: Proceedings of the Fourth International Symposium on Uncertainty Modeling and Analysis (ISUMA 2003), pp. 386–391. IEEE Press, New York (2003)
10. Varghese, B., Poojari, C.A.: Genetic algorithm based technique for solving chance constrained problems arising in risk management. Technical Report, Carisma (2004)

Characteristic Analysis and Management of Human Resource in EPC Project

Wenhai Luo, Yanfang Zhu, and Tao Hong

School of Civil and Architectural Engineering, Nanchang Institute of Technology, No. 289,
Tianxiang Road, State HIGH-Tech Industry Development Zone, Nanchang, 330099, China
DSYTX@126.com

Abstract. Based on the theory of project management and human resources, combining with the practical construction situations in China, using theoretical analysis and comparative analysis, four important characteristics are extracted in engineering project using EPC mode, that is, mobility, duality, commodity & humanity and collaboration. Furthermore, it is analyzed the origins and implications of above characteristics in depth, and some management methods are presented with the guidance in connection with them in theory.

Keywords: EPC, human resource management, characteristic analysis.

1 Introduction

It is not accidental that EPC mode is widely used in international engineering area, but it is the result which shows EPC is more prominent comparing with CM, DB and other modes. Although EPC starts a bit late in China, it expands rapidly. Generally, the engineering projects utilizing EPC mode refers to larger scale, long period, complex techniques and professional skills, in order to reduce the risks in engineering construction, the owner needs to allocate sophisticated engineering professors to enhance the supervision of EPC prime contractor[1], the uniqueness of using HR in EPC mode is shown. In this case, some analysis of the characteristics of HR in EPC are shown in details, and some guiding management methods about different characteristics are represented.

2 The Characteristics of Human Resource in EPC Mode

2.1 Mobility

In EPC mode, the sources of HR can be separated into internal and external human resource. The mobility of internal human resource is relatively smaller, whereas it would be mobilized accompanying the changes of the project, including regional alterations and different personnel allocations in diverse projects. As for external human resource, the mobility is quite large due to that most are provisional employees

H. Sun et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 329, DOI: 10.1007/978-3-319-12286-1_4

or even peasant-workers, and different requirements on the amount of workers and techniques during each construction period, and even the transformation of seasons and vocations. Normally, in order to control the cost, an EPC engineering company with several construction projects under the same commencement date would maximize the utilization of HR which is the reason why personnel is in charge of several projects or enters into another project right after completing the former one. All these cases determine that HR in EPC project has to be kept for an extended term and hardly has a break.

2.2 Duality

The structure of organization of HR in EPC can be depicted in matrix (Fig. 1).

The above chart shows clearly that the member in EPC human resource is not just a part of personnel in a company but also a team member in a specific engineering project that is restrained by both company and project department, and hence undertaking the responsibilities and roles in two fold ways. It explicitly takes the project as orientation and indicates that the project management team assumes the coordination work. Besides, even though the human resources come from different functional departments that enable them to support each other technically, each certified staff still belongs to their own functional department and needs to return after the project finished. This is the duality of HR in EPC.

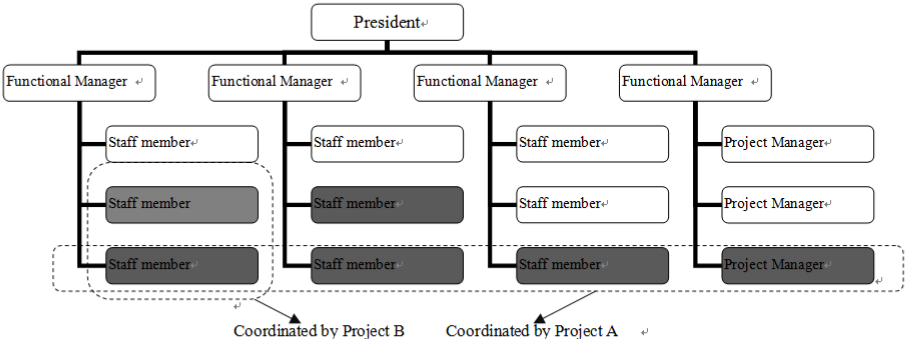


Fig. 1. Organization Chart of HR in EPC

2.3 Commodity and Humanity

HR is a general term of all manual and mental workers who promote social and economic development, create material and spiritual wealth. it is also a part of the labor force resources, but its purpose is more. In the principle of Marxism, labor force deemed as a type of commodity like anything else, has the same two attributes which are value and use value. Therefore, the commodity of labor force resources determines the viable human resources. However, on the other hand, humanity is a special type of commodity with social attribute, requiring for humanistic and life caring, and needs to be respected, cherished and concerned, which shows the humanity in HR.

2.4 Collaboration

In traditional ways of engineering construction (Fig. 2), the procurement process can't be carried out before the accomplishment of the project design. Because of the slow process in procurement, the start of construction also has to wait until all purchases finish which normally delays of the construction period, leading to unnecessary increase in cost.

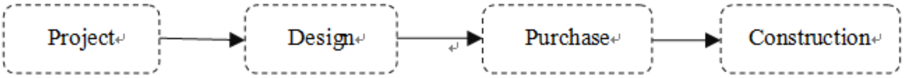


Fig. 2. Traditional construction project flowcharts

By contrast, it is doable to projects in the EPC designed-build engineering mode (Fig.3) to integrate design, purchase and construction in the same period. In other words, equipment procurement can be carried out during designing process which effectively reduces the construction period as well as the cost. Additionally, in EPC DB mode, the Construction Method Statement can be optimized, thus, engineering quality can be enhanced and resources can be saved during constructing.

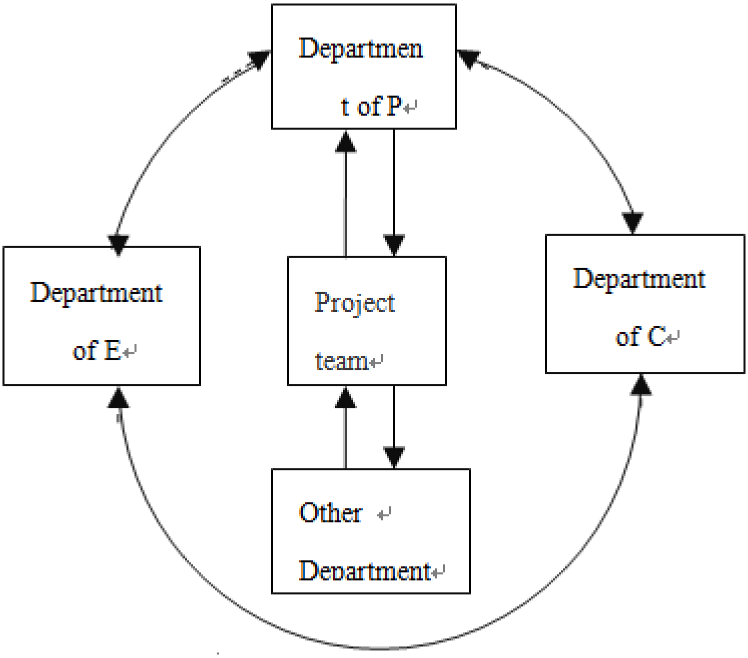


Fig. 3. EPC project flowchart(E: Design P: Purchase C: Construction)

According to the comparison of two flow charts above, design department, in traditional construction mode, is not necessary to contact much with purchasing and construction department after handing over the design. However, in EPC DB mode, construction department is able to communicate with purchasing and construction department once they find any designs or problems of products when constructing to make relevant alterations. Similarly, design department can also make some proposals when realizing the purchasing department fails to understand the intention of designing or finding the construction department modifies the design to shorten the time without permission. Therefore, compared with customary project, the three departments in EPC are more collaborative ,which has become the essential requirement for EPC to go well and finish on time.

3 Management towards Different Characteristics in HR

3.1 Management Aiming at Mobility

(1) Defining responsibility, allocating roles

In each EPC project, it is necessary to establish the standard job description, including work titles, activities, procedures, physical circumstance, social environment and conditions of employment, etc., and it is also a necessity to possess detailed office-holding specifications to state the physical and psychological requirements to the employees in certain positions. Job description and office-holding specifications are the utmost approaches to response characteristic change in a team. It can guarantee that every member in a new group has a clear obligation and distribution and there would be a slight influence if some members flow, and more importantly, it can provide unambiguous information for new recruits or substitutes.

(2) Formulating reasonable allowance and subsidy system

In EPC projects , employees travel frequently. In addition to the basic wage salaries, allowances and subsidies section also should be included. To achieve the purposes that rewarding according to labor expended, increasing revenue, inspiring energy and soothing. In EPC engineering company, in addition to the basic reimbursement system, it must also adjust the principles and standards of issuance of subsidies and allowances which should reasonably reflect the different levels in remote areas, and also increase the standard of allowances according to changes in local environment and social developments. And also the establishment of the bonus system based on the profits of each project proceeds can be taken into account, giving exceptional rewards for employees who make outstanding contributions and work in the first line for long-term .

(3) Creating the conditions to meet the spiritual and cultural life needs of employees

In the management of EPC HR, it is particularly essential to enrich employees' spiritual and cultural life as well as satisfy their needs. Firstly, EPC engineering company needs to carry out the learning activities in a planned, focused, targeted organization to guide and support the staffs. Secondly, it are required to insist people oriented and organize wholesome spiritual and cultural activities regularly. Owing to the employees still working in the construction site in holidays, company is advised to organize relevant condolence and celebration. And for those who have family difficulties or are unable to return to resolve emergencies due to work, some help is needed as well. Thirdly, it is necessary to implement the medical security, improve staff welfare and ensure the health of employees, so that employees can remain full of energy all the time.

3.2 Management Aiming at Duality

When discussing the methods in the management of duality in EPC HR, we should commence from two organizations and analyze their management in contents and goals.

(1) HR management department in EPC engineering company-Functional Management

Operative management, in another word, is the management of duty and performance. One elemental feature of operative management is to turn the repeatable production operations into a series of standard and ordered tasks and assign to specific performers. The final purpose is to keep the company developing continuously, effectively and healthily. The HR Management Department in EPC engineering company operates centered on functional management aiming at whole staffs in the company and the method to achieve management goals is through the work of the specific functional departments.

(2) EPC HR management --performance management

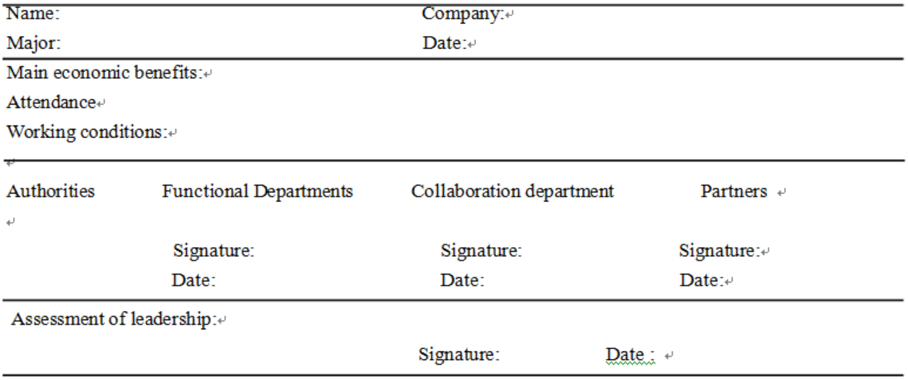
Performance is the achievement of works [2], and is the unity of behavior, direct outcomes and final outcomes of individuals or groups [3]. Performance management is the continuous circulation of planning, communication, assessment, application of the results, and enhancement associating with each rank of employers and employees aiming at achieving the same organizational target. The purposes of performance management also include enhancing performances of people, departments and organizations. EPC HR management is not only anthropological resource management of EPC projects, but also that of project teams. It treats working outcomes and pecuniary profits as priorities and is an overall process of construction and management of teams with performance assessment throughout.

Common performance assessments of EPC engineering projects include two main types: the first one is evaluation of outcome which assess employee mainly from top to bottom to clarify working capacities and devotions of each employees, and to develop skills and personal qualities of employees as well as enhance solidarity, fighting will, competitiveness and flexibility of enterprises. It is for long-term assessments and designs. The assessments table can be designed as follows.

Table 1. Performance evaluation table

Assessment content Basic information	Name:	Gender:	Post:	Major:
Political literacy				
Professional technical ability				
Knowledge update				
Public relations				
Comment:				
			Signature:	Date:

Table 2. Effectiveness evaluation table



Name:		Company:	
Major:		Date:	
Main economic benefits: Attendance Working conditions:			
Authorities	Functional Departments	Collaboration department	Partners
	Signature: Date:	Signature: Date:	Signature: Date:
Assessment of leadership:			
		Signature:	Date：

The second one is profits for assessment. This assessment is based on economic objectives, targets as project teams' members according to their practical workloads. The alterations in salary can be based on this assessment. The assessments table can be designed as follows.

3.3 Management in Commodity and Humanity

(1) Commodity Management

We need to strengthen our marketing concepts when dealing with EPC HR. Besides, when allocating human resources, project teams should compare expenditure cost between internal HR and external HR, and should be possible to make use of external human resources which is the cheap, reliable human resources surrounding the construction site, and internal human resources saved can be used under other conditions of lack of external human resources projects. Therefore, it is necessary to investigate situations of human resources before EPC projects start. Investigation details are as follows:(1) distributions of HR around sites of projects, (2) conditions of each type of current HR and numbers of available contract workers or temporary workers, (3) the possibility of expanding worker recruitment based on current situation. The research before project starts is the only way to assure the cost in investigation of HR can be in line with the actual situation of the market.

(2) Humanity management

Chinese characteristics of modern human spirit is what we are advocating a "people-oriented".

The core of humanity is human, which is what we and the customary Chinese culture advocate--human oriented. Human oriented management of EPC engineering projects should take following statements into consideration:(1) appropriate stress for employees to move forward, (2) reasonable rules to regulate certain behaviors, (3) proper awards to encourage competition among employees, (4) lawful means to guarantee employees’ rights and interests legally; (5) suitable working environments.

3.4 Management of Collaboration

Economic globalization makes knowledge, information, internet and innovation become platform for running projects in the current environment between competition and cooperation [4][5].In order to enhance cooperation and coordination among design department, purchase department and construction department, efficient network for information exchange needs to be built-in the first place, which accelerates the process of communication among each departments, shares information and resources, makes communication and information exchange process convenient between employers and employees, project teams and the outside world, and among departments, and also enhances efficiency and reduces costs. In addition, democratic centralism is expected to be utilized to guarantee ascendancy and predominance of the project management team which enables the team to deal with a variety of problems associating with projects and renders preventive measures and solutions as soon as possible, keep projects running efficiently and improve success rates of projects as a whole.

4 Conclusion

According to the characteristics of EPC human resources, based on years of engineering practice, special management methods are proposed, the method of EPC human resources management established which can be used by all EPC project, and a suitable method for the new EPC project is provided which can be used for human resources management in the shortest possible time, which combining four characteristics of the EPC human resources four of scientific management methods, grasping the core content of management, mastering the main content of management, weakening the other auxiliary conditions. It shows that the management method is scientific, accurate, intuitive and practical. So the method can be mastered and operated by many kinds of department managers, and it has a positive effect for the whole project which demanding for economy, progress under control and completion on time.

Acknowledgments. This work is supported by the Young Foundation Plan of Nanchang Institute of Technology under Grant(No.2012SK008).

References

1. Li, Z., He, J.: Select About EPC project management model and study the impact on the efficient operation of the project. Hubei: Science&Technology Progress and Policy 27(19), 7–10 (2010)
2. Armstong, M., Baronl, A.: Performance Management, pp. 15–52. The Cromwell Press (1998)
3. Zhang, B.: Human Resource Management, pp. 88–105. Nanjing University Press (2000)
4. Pinto, J.K.: Project management 2002. Research Technology Management 45(2), 22–37 (2002)
5. Jiang, J.J., Klein, G., Chen, H.-G.: The relative influence of IS project implementation policies and project leadership on eventual outcomes. Project Management Journal 32(3), 49–55 (2001)
6. Yang, H.: The study of human resource allocation in project management. Journal of Architectural Education in Institutions of Higher Learning 17(3), 61–64 (2008)
7. Armstong, M., Baronl, A.: Performance Management, pp. 15–52. The Cromwell Press (1998)
8. Du, C.: The study of the mode of projects human resource management. Harbin Engineering University's department of engineering management (2005)
9. Thamhain, H.J.: Leading Technology-Based Project Teams. Engineering Management Journal 16(2), 35–38 (2004)
10. Wang, Z., Yang, G.: Project Management-Principles and Case, 2nd edn., pp. 42–68. China Water Power Press (2009)

Hybrid Particle Swarm Optimization with Bat Algorithm

Tien-Szu Pan[1], Thi-Kien Dao[1], Trong-The Nguyen[2], and Shu-Chuan Chu[3]

[1] Department of Electronics Engineering,
National Kaohsiung University of Applied Sciences, Taiwan
[2] Faculty of Information Technology, Haiphong Private University, Vietnam
[3] School of Computer Science, Engineering and Mathematics,
Flinders University, Australia
vnthe@hpu.edu.vn

Abstract. In this paper, a communication strategy for hybrid Particle Swarm Optimization (PSO) with Bat Algorithm (BA) is proposed for solving numerical optimization problems. In this work, several worst individuals of particles in PSO will be replaced with the best individuals in BA after running some fixed iterations, and on the contrary, the poorer individuals of BA will be replaced with the finest particles of PSO. The communicating strategy provides the information flow for the particles in PSO to communicate with the bats in BA. Six benchmark functions are used to test the behavior of the convergence, the accuracy, and the speed of the approached method. The results show that the proposed scheme increases the convergence and accuracy more than BA and PSO up to 3% and 47% respectively.

Keywords: Hybrid Particle Swarm Optimization with Bat Algorithm, Particle Swarm Optimization Algorithm, Bat Algorithm Optimizations, Swarm Intelligence.

1 Introduction

Computational intelligence algorithms have been prosperously used to solve optimization problems in engineering, economic, and management fields for recently years. For example, genetic algorithms (GA) have been used prosperously in various applications, including engineering, the budgetary and the security area [1-3]. Particle swarm optimization (PSO) techniques have fortunately been used to forecast the exchange rates, to optimize related multiple interference cancellations [4-6], to construct the portfolios of stock, human perception [3, 7, 8]. Ant colony optimization (ACO) techniques have successfully been used to solve the routing problem of networks, the secure watermarking [9, 10]. Artificial bee colony (ABC) techniques have successfully been used to solve the lot-streaming flow shop scheduling problem [11]. Cat swarm optimization (CSO) [12] techniques have successfully been used to discover proper positions for information hiding [13].

Communication between two algorithms is to take the advantage of the strength points of each type of algorithms. The idea of this paper is based on communication

H. Sun et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 329, DOI: 10.1007/978-3-319-12286-1_5

strategies in parallel processing for swarm intelligent algorithms. Information between populations is only exchanged when the communication strategy is triggered. The parallel strategies simply share the computation load over several processors. The sum of the computation time for all processors can be reduced compared with the single processor works on the same optimum problem. In this paper, the concepts of parallel processing and communication strategy are applied to hybrid Particle Swarm Optimization with Bat Algorithm is presented. In this new method, the several poorer particles in PSO will be replaced with best artificial bats in Bat algorithm after running some fixed iterations and on the contrary, the poorer individuals of BA will be replaced with the better particles of PSO.

The rest of this paper is organized as follows: a brief review of PSO and BA is given in session 2; our analysis and designs for the hybrid PSO-BA is presented in session 3; a series of experimental results and the comparison between PSO, BA and hybrid PSO-BA are discussed in session 4; finally, conclusion is summarized in session 5.

2 Related Work

Particle swarm optimization (PSO) is a heuristic global optimization algorithm, based on the research of bird and fish flock in movement behavior, proposed by Kennedy, J. Eberhart, R. and Shi [14, 15]. The particles are randomly initialized and then freely fly across the multi-dimensional search space. While they are flying, its velocity and position are updated based on its own best experience and also of entire population. The updating policy will cause the particle swarm to move toward a region with a higher object value. The position of each particle is equivalent to a candidate solution of a problem. The particle moves according to an adjusted velocity, which is based on that particle's experience and the experience of its companions.

The original particle swarm optimization algorithm can be expressed as follows:

$$V_i^{t+1} = V_i^t + C_1 \times r_1(P_i^t - X_i^t) + C_2 \times r_2(G^t - X_i^t) \tag{1}$$

where V_i^t is the velocity of the *i-th* particle at the *t-th* iteration, C_1 and C_2 are factors of the speed control, r_1 and r_2 are random variables such that $0 \leq r_1, r_2 \leq 1$, P_i^t is the best previous position of the *i-th* particle at the *t-th* iteration, G^t is the *best* position amongst all the particles, from the first iteration to the t-$^{\text{th}}$ iteration, and X_i^t is the *i-th* particle for the *t-th* iteration.

$$X_i^{t+1} = X_i^t + V_i^{t+1}, i = 0,1,..N - 1 \tag{2}$$

where N is the particle size, $-V_{\max} \leq V^{t+1} \leq V_{\max}$ ($V_{\max}$ is the maximum velocity).

A modified version of the particle swarm optimizer [15] and an adaption using the inertia weight which is a parameter for controlling the dynamics of flying of the modified particle swarm [16], have also been presented. The latter version of the modified particle swarm optimizer can be expressed as equation (3)

$$V_i^{t+1} = W^t \times V_i^t + C_1 \times r_1(P_i^t - X_i^t) + C_2 \times r_2(G^t - X_i^t) \tag{3}$$

$$X_i^{t+1} = X_i^t + V_i^{t+1}, \quad i = 0,1,..N-1 \tag{4}$$

where W^t is the inertia weight at the t-th iteration.

Moreover, the Bat Algorithm (BA) is a new optimization algorithm, proposed by Xin-SheYang, based on swarm intelligence and the inspiration from observing the bats [17]. BA simulates parts of the echolocation characteristics of the micro-bat in the simplicity way. Three major characteristics of the micro-bat are employed to construct the basic structure of BA. The idealized rules in this method are listed as follows: The echolocation to detect the prey is utilized for all bats, but not all species of the bat do the same thing. However, the micro-bat, one of species of the bat is a famous example of extensively using the echolocation. Hence, the first characteristic is the echolocation behavior. The second characteristic is the frequency. The frequency is sent by the micro-bat with fixed frequency f_{min} and with a variable wavelength λ. The loudness A_0 is used to search for prey. The other characteristic of them are listed as follows:

1. Bats fly randomly with velocity v_i at position x_i. They can adjust the wavelength (or frequency) of their emitted pulses and adjust the rate of pulse emission *r* from 0 to1, depending on the proximity of their target;
2. There are many ways to adjust the loudness. For simplicity, the loudness is assumed to be varied from a positive large A_0 to a minimum constant value, which is denoted by Amin.

The movement of the virtual bat is simulated by equation (5) – equation (7):

$$f_i = f_{min} + (f_{max} - f_{min}) * \beta \tag{5}$$

$$v_i^t = v_i^{t-1} + (x_i^{t-1} - x_{best}) * f_i \tag{6}$$

$$x_i^t = x_i^{t-1} + \mathrm{v}_\mathrm{i}^\mathrm{t} \tag{7}$$

where f is the frequency used by the bat seeking for its prey, f_{min} and f_{max}, represent the minimum and maximum value, respectively. *xi* denotes the location of the *i-th* bat in the solution space, *vi* represents the velocity of the bat, *t* indicates the current iteration, β is a random vector, which is drawn from a uniform distribution, and β ∈[0, 1], and x_{best} indicates the global near best solution found so far over the whole population.

In addition, the rate of the pulse emission from the bat is also taken to be one of the roles in the process. The micro-bat emits the echo and adjusts the wavelength depending on the proximity of their target. The pulse emission rate is denoted by the symbol r_i, and r_i ∈[0,1], where the suffix *i* indicates the *i-th* bat. In every iteration, a random number is generated and is compared with r_i. If the random number is greater than r_i, a local search strategy, namely, random walk, is detonated. A new solution for the bat is generated by equation (8):

$$x_{new} = x_{old} + \varepsilon A^t \tag{8}$$

where ε is a random number and $\varepsilon \in[-1,1]$, and the average loudness of all bats is represented at the current time step *t*. After updating the positions of the bats, the loudness A_i and the pulse emission rate r_i are also updated only whenever the global near best solution is updated and the random generated number is smaller than A_i. The update of A_i and r_i are operated by equation (9) and equation (10):

$$A_i^{t+1} = \alpha A_i^t \tag{9}$$

$$r_i^{t+1} = r_i^0[1 - e^{-\gamma t}] \tag{10}$$

where α and γ are constants. In Yang's experiments, $\alpha = \gamma = 0.9$ is used for the simplicity.

The process of BA is depicted as follows:

Step 1. Initialize the bat population, the pulse rates, the loudness, and define the pulse frequency

Step 2. Update the velocities to update the location of the bats, and decide whether detonate the random walk process.

Step 3. Rank the bats according to their fitness value, find the current best solution found so far, and then update the loudness and the emission rate.

Step 4. Check the termination condition to decide whether go back to step 2 or end the process and output the result.

3 The Proposed Hybrid PSO with BA

Hybrid optimization algorithm is structured by communication strategies between two algorithms in this paper. This idea is based on replacing the weaker individuals according to fitness evaluation of one algorithm with stronger individuals from other algorithm in parallel processing for swarm intelligent algorithms. Several groups in a parallel structure of hybrid algorithm are created by dividing the population into subpopulations. Each of the subpopulation evolves independently in regular iterations. They only exchange information between populations when the communication strategy is triggered. It results in taking advantage of the individual strengths of each type of algorithm. The replacement of weaker individuals in running algorithms will be achieved so on to get the benefit of the cooperation.

Hybrid Particle Swarm Optimization with Bat algorithm (hybrid PSO-BA) is designed based on original PSO and Bat algorithm. Each algorithm evolves by optimization independently, i.e. the PSO has its own individuals and the better solution to replace the worst artificial bats of BA. In contrast, the better artificial bats of BA are to replace the poorer individuals of PSO after running some fixed iterations. The total iteration contains R times of communication, where $R = \{R_1, 2R_1, 3R_1, \ldots\}$.

Let N be the number of population of hybrid PSO-BA, and N_1, N_2 be the number of population of PSO and BA respectively, where N_1 and N_2 are set to be $N/2$. If $t \cap R \neq \varphi$, k agents with the top k fitness in N_1 will be copied to N_2 to replace the same number of individuals with the worst fitness, where t denotes the current iteration count, R_1 and k are the predefined constants.

The diagram of the hybrid PSO-BA with communication strategy is shown in figure 1

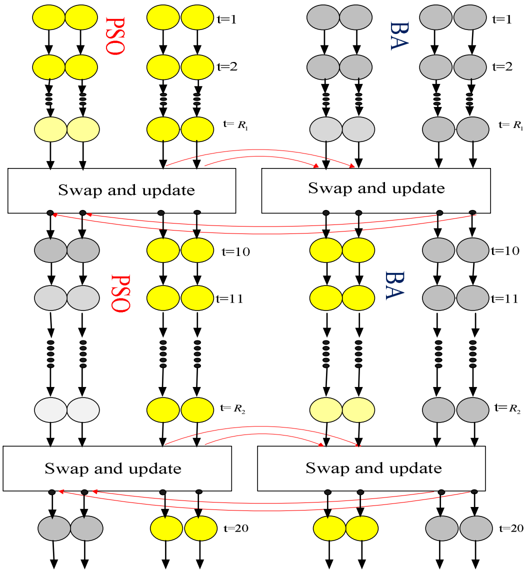


Fig. 1. The diagram of hybrid PSO-BA with a communication strategy

1. **Initialization:** Generate populations for both PSO and BA. Each population is initialized by BA or by PSO independently. Defined the iteration set R for executing the communication strategy. The N_1, N_2 are the number of particles and artificial agents in solutions S_{ij}^T and X_{ij}^T for populations of PSO and BA respectively, $i = 0, 1, \ldots, N_1 - 1$, $j = 0, 1,..D$. D is dimension of solutions and t is current iteration number. Set $t = 1$.

2. **Evaluation:** Evaluate the value of $f_1(S_{ij}^T)$, $f_2(X_{ij}^T)$ for both PSO and BA in each population. The evolvement of the populations is executed independently by both PSO and BA.

3. **Update:** Update the velocity and the positions of PSO using equation (1), and (2). Update the location and velocity of Bat in the best fitness value, which are found by the bat using equation (6), (7).

4. **Communication Strategy:** Migrate the best artificial bats among all the individuals of BA's population, copy k bats with the top k fitness in N_1 replace the poorer particles in N_2 of PSO's population and update for each population every R_1 iterations.

5. **Termination:** Repeat step 2 to step 5 until the predefined value of the function is achieved or the maximum number of iterations has been reached. Record the best value of the function $f(S^t)$ and the best particle position among all the particles S^t. Record the best value of the function $f(X^t)$ and the best location among all the bats X^t.

4 Experimental Results

This section presents simulation results and compares the hybrid PSO-BA with the primary PSO, and original BA, both in terms of solution quality, convergence capability, and the accuracy. The execution times in the number of function evaluations are also taken. Six benchmark functions are used to test the accuracy and the convergence of hybrid PSO-BA.

All the benchmark functions for experimenting are averaged over different random seeds with 10 runs. Let $S = \{s_1, s_2 ,..., s_m\}$, $X = \{x_1, x_2 ,..., x_m\}$ be the m-dimensional real-value vectors for PSO and BA respectively. The benchmark functions are Ackley, Griewank, Quadric, Rastrigin, Rosenbrock and Spherical. The equation numbers (11) to (16). The goal of the optimization is to minimize the outcome for all benchmarks. The population size of hybrid PSO-BA, primary PSO and original BA are set to 20 (N=*20*) for all the algorithms in the experiments. The detail of parameter settings of PSO can be found in [14], and setting of BA can be found in [17].

$$f_1(x) = 20 + e - 20e^{-0.2\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}}} - e^{\frac{\sum_{j=1}^{n} \cos(2\pi x_i)}{n}} \tag{11}$$

$$f_2(x) = 1 + \sum_{i=1}^{N} \frac{x_i^2}{4000} + \prod_{i=1}^{N} cos\frac{x_i}{\sqrt{i}} \tag{12}$$

$$f_3(x) = \sum_{i=1}^{n}(\sum_{k=1}^{i} x_i) \tag{13}$$

$$f_4(x) = \sum_{i=1}^{N}[10 + x_i^2 - 10cos2\pi x_i \tag{14}$$

$$f_5(x) = \sum_{i=1}^{n-1}(100(x_{i-1} - x_i^2)^2 + (1 - x_i)^2 \tag{15}$$

$$f_6(x) = \sum_{i=1}^{N} x_i^2 \tag{16}$$

The initial range and the total iteration number for all test functions are listed in Table1.

Table 1. The initial range and the total iteration of test standard functions

Function		Initial range	Total iteration
		$[x_{min}, x_{max}]$	
Ackley	$f_1(x)$	[-100,100]	200
Griewangk	$f_2(x)$	[5.12,5.12]	200
Quadric	$f_3(x)$	[-100,100]	200
Rastrigin	$f_4(x)$	[-30,30]	200
Rosenbrock	$f_5(x)$	[-100,100]	200
Spherical	$f_6(x)$	[-100,100]	200

The optimization for all of these test functions is to minimize the outcome. The parameters setting for hybrid PSO-BA with primary PSO side are the initial inertia weight $W = (0.9 - 07 * rand)$, coefficients of learning factors c_1=-2 and c_2=2 in PSO, the total population size N_1 = 10 and the dimension of the solution space M = *10*, and with original BA side are the initial loudness $A_i^0 = 0.25$, pulse rate $r_i^0 = 0.5$ the total population size N_1 = 10 and the dimension of the solution space M = *10*, frequency minimum f_{min} = *the lowest of initial range function* and frequency minimum f_{max} = *the highest of initial range function.* The proposed scheme is executed for 10 runs and each run contains 200 iterations. The final result is obtained by taking the average of the outcomes from all runs. These results also are compared with the primary PSO and original BA respectively.

Table 2 compares the quality of optimizing performance and time running for numerical problem optimization between hybrid PSO-BA and PSO. It is clearly seen that, almost these cases of testing benchmark functions for hybrid PSO-BA are better than PSO in terms of convergence and accuracy. It is special case with test function of Rosenbrock, $f_5(x)$ has the mean of value function minimum of total seeds of 10 runs is 1.02E+09 for hybrid PSO-BA performance evaluation, but, for original PSO is 2.90E+09, reaches at 48% improvement of convergence. The average performance evaluation value of six benchmark functions is 1.70E+08 for hybrid PSO-BA and 4.83E+08 for original PSO, gets at 47% improvement of accuracy. However, all benchmark functions for average time consuming of hybrid BA-BA are longer than that in original PSO, for the reasons, the hybrid algorithm must perform mutation and update operations.

Table 2. The comparison between hybrid PSO-BA and origianl PSO in terms of quality performance evaluation and speed

Function	Performance evaluation		Time running evaluation (seconds)	
	PSO	***Hybrid PSO-BA***	***PSO***	***Hybrid PSO-BA***
$f_1(x)$	1.96E+01	1.85E+01	0.079	0.134
$f_2(x)$	1.92E+00	1.81E+00	0.086	0.139
$f_3(x)$	4.46E+03	2.62E+03	0.109	0.230
$f_4(x)$	1.23E+02	1.11E+02	0.080	0.148
$f_5(x)$	2.90E+09	1.02E+09	0.081	0.159
$f_6(x)$	1.62E+04	7.13E+03	0.064	0.121
Average value	**4.83E+08**	**1.70E+08**	**0.33**	**0.49**

Figure 2 shows the experimental results of six benchmark functions in running repeatedly same iteration of 200 in random seeds of 10 runs. It clearly can be seen that the most cases of curves of hybrid PSO-BA (solid red line) are more convergence that its of PSO (doted blue line).

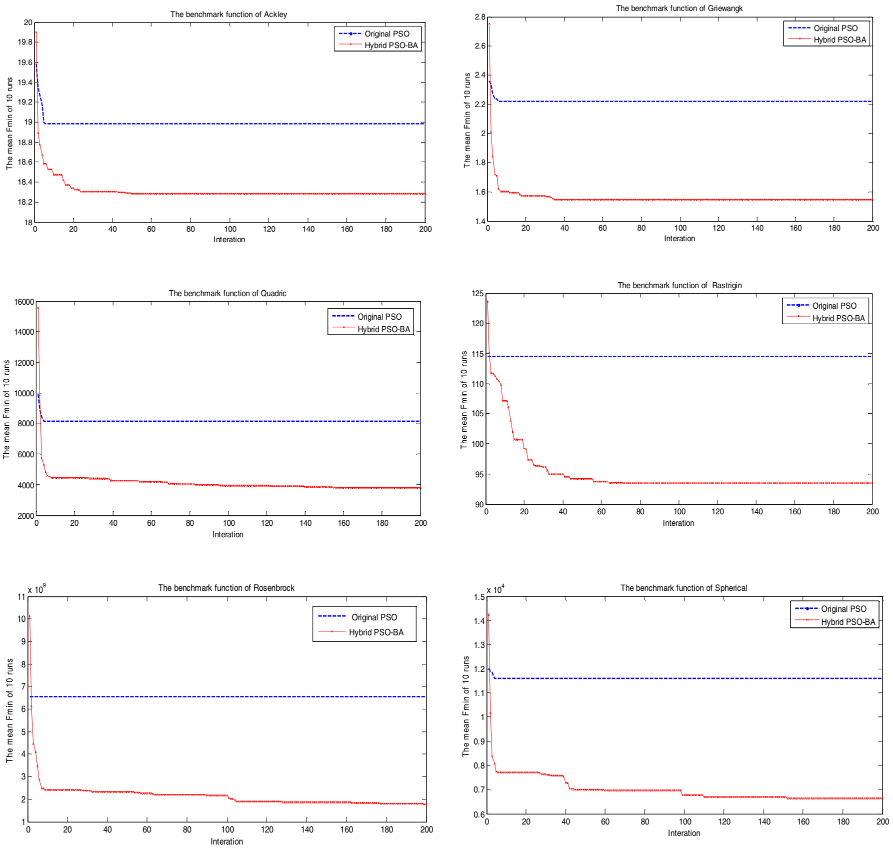


Fig. 2. The mean of function minimum curves in comparing Hybrid PSO-BA and original PSO algorithms for function of Ackley, Griewank, Quadric, Rastrigin, Rosenbrock and Spherical

Table 3 compares the quality of performance and time running for numerical problem optimization between Hybrid BA-PSO and original BA. It is clearly seen that, almost these cases of testing benchmark functions for Hybrid BA-PSO are more convergence than original BA.

Average value of all benchmark functions for hybrid BA- PSO is 2.32E+07 in performance evaluation, but this figure is 2.30E+07 for original BA, reaches at 3% improvement of accuracy. However, average times consuming of all benchmark functions for hybrid BA-PSO is longer taken than original BA. For this result, the reason is the hybrid algorithm must perform mutation and update operations.

Table 3. The comparison between hybrid BA-PSO and origianl BA in terms of quality performance evaluation and speed

Function	Performance evaluation		Time running evaluation (seconds)	
	BA	***Hybrid BA-PSO***	***BA***	***Hybrid BA-PSO***
$f_1(x)$	1.84E+01	1.65E+01	0.087	0.104
$f_2(x)$	7.37E-01	7.34E-01	0.094	0.119
$f_3(x)$	2.59E+03	2.05E+03	0.120	0.180
$f_4(x)$	4.67E+01	4.60E+01	0.087	0.098
$f_5(x)$	1.38E+08	1.38E+08	0.089	0.109
$f_6(x)$	2.65E+03	2.47E+03	0.071	0.071
Average value	**2.35E+07**	**2.30E+07**	**0.101**	**0.124**

Figure 3 shows the experimental results of six benchmark functions in running 10 seeds output with the same iteration of 200. It clearly can be seen that the most cases of curves of hybrid BA-PSO (solid red line) are more convergence that its of BA (doted blue line).

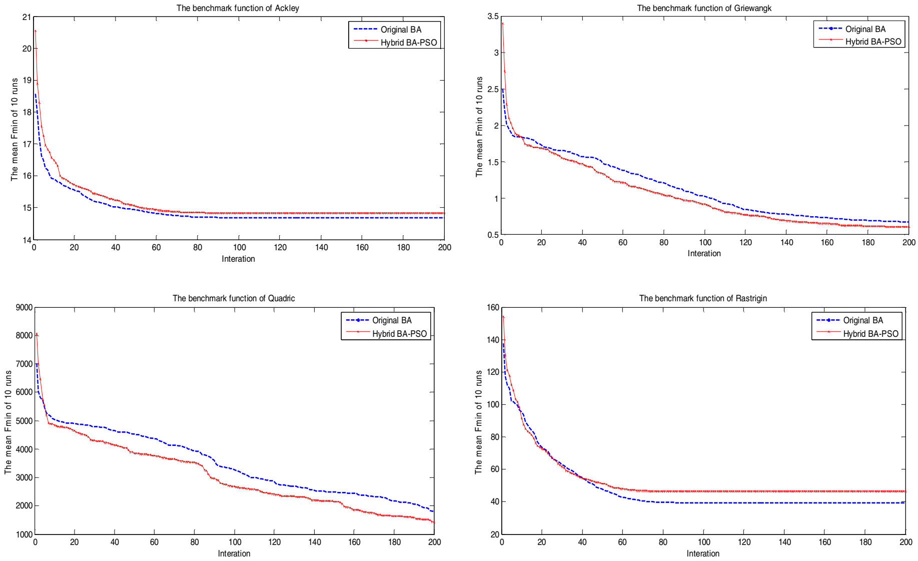


Fig. 3. The mean of function minimum curves in comparing hybrid BA-PSO and BA algorithms for function of Ackley, Griewank, Quadric, Rastrigin, Rosenbrock and Spherical

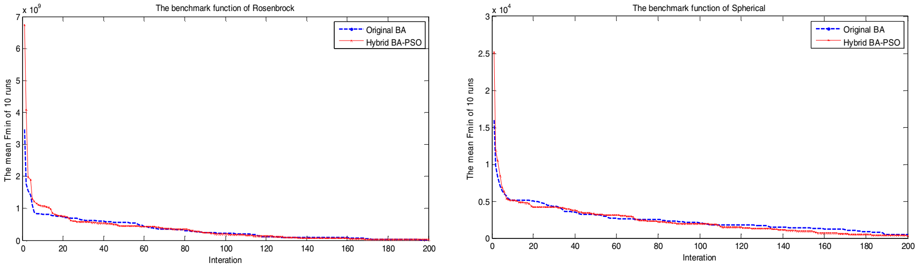


Fig. 3. (*continued*)

5 Conclusion

This paper, a novel proposed optimization scheme was presented, namely hybrid PSO-BA (hybrid Particle Swarm Optimization with Bat Algorithm). The implementation of hybrid for optimization algorithms could have important significance for taking advantages of the power of each algorithm and achieving cooperation of optimization algorithms. In the new proposed algorithm, the several worse individuals in PSO are replaced with the best artificial bats in BA algorithm after running some fixed iterations, and on the contrary, the poorer bats of BA are replaced with the better particles of PSO.

The proposed communication strategy provides the information flow for the particles to communicate in PSO with the bats in BA. The performance of hybrid PSO-BA algorithm is better than both original PSO and BA in terms of convergence and accuracy. The results the proposed algorithm on a set of various test problems show that hybrid PSO-BA increases the convergence and accuracy more than original PSO and original BA is up to 47 % is at 3% on finding the near best solution improvement.

References

1. Srinivas, M., Patnaik, L.M.: Genetic algorithms: a survey. Computer 27(6), 17–26 (1994)
2. Yang, B., Niu, X., Wang, S.: A Secure Steganography Method based on Genetic Algorithm. Journal of Information Hiding and Multimedia Signal Processing 1(1), 8 (2010)
3. Ruiz-Torrubiano, R., Suarez, A.: Hybrid Approaches and Dimensionality Reduction for Portfolio Selection with Cardinality Constraints. IEEE Computational Intelligence Magazine 5(2), 92–107 (2010)
4. Chen, S.-M., Chien, C.-Y.: Solving the traveling salesman problem based on the genetic simulated annealing ant colony system with particle swarm optimization techniques. Expert Systems with Applications 38(12), 14439–14450 (2011)
5. Shyr, W.-J., Hsu, C.-H., Kuo, K.-H.: Optimizing Multiple Interference Cancellations of Linear Phase Array Based on Particle Swarm Optimization. Journal of Information Hiding and Multimedia Signal Processing 1(4), 292–300 (2010)

6. Chen, S.-M., Kao, P.-Y.: TAIEX forecasting based on fuzzy time series, particle swarm optimization techniques and support vector machines. Information Sciences 247(0), 62–71 (2013)
7. Jui-Fang, C., Shu-Wei, H.: The Construction of Stock_s Portfolios by Using Particle Swarm Optimization, p. 390
8. Puranik, P.B.P., Abraham, A., Palsodkar, P., Deshmukh, A.: Human Perception-based Color Image Segmentation Using Comprehensive Learning Particle Swarm Optimization. Journal of Information Hiding and Multimedia Signal Processing 2(3), 227–235 (2011)
9. Pinto, P.C., Nagele, A., Dejori, M., Runkler, T.A., Sousa, J.M.C.: Using a Local Discovery Ant Algorithm for Bayesian Network Structure Learning. IEEE Transactions on Evolutionary Computation 13(4), 767–779 (2009)
10. Chouinard, J.-Y., Loukhaoukha, K., Taieb, M.H.: Optimal Image Watermarking Algorithm Based on LWT-SVD via Multi-objective Ant Colony Optimization. Journal of Information Hiding and Multimedia Signal Processing 2(4), 303–319 (2011)
11. Pan, Q.-K., Tasgetiren, M.F., Suganthan, P.N., Chua, T.J.: A discrete artificial bee colony algorithm for the lot-streaming flow shop scheduling problem. Inf. Sci. 181(12), 2455–2468 (2011)
12. Chu, S.-C., Tsai, P.-W.: Computational Intelligence Based on the Behavior of Cats. International Journal of Innovative Computing, Information and Control 3(1), 8 (2006)
13. Wang, Z.-H., Chang, C.-C., Li, M.-C.: Optimizing least-significant-bit substitution using cat swarm optimization strategy. Inf. Sci. 192, 98–108 (2012)
14. Kennedy, J., Eberhart, R.: Particle swarm optimization, vol. 4, pp. 1942–1948 (1995)
15. Yuhui, S., Eberhart, R.: A modified particle swarm optimizer, pp. 69–73 (1998)
16. Yuhui, S., Eberhart, R.C.: Empirical study of particle swarm optimization, vol. 3, p. 1950 (1999)
17. Yang, X.-S.: A New Metaheuristic Bat-Inspired Algorithm. In: González, J.R., Pelta, D.A., Cruz, C., Terrazas, G., Krasnogor, N. (eds.) NICSO 2010. SCI, vol. 284, pp. 65–74. Springer, Heidelberg (2010)

Interaction Artificial Bee Colony Based Load Balance Method in Cloud Computing

Jeng-Shyang Pan[1], Haibin Wang[2], Hongnan Zhao[1], and Linlin Tang[1]

[1] Harbin Institute of Technology Shenzhen Graduate School,
Shenzhen, China
[2] ShenZhen State High-Tech Industrial Innovation Centre,
Shenzhen, China
{jengshyangpan,linlintang2009}@gmail.com, wanghb@szcxzx.net,
zhaohongnan1234@163.com

Abstract. Rapidly development of the cloud computing and Internet makes load balance technique become more and more significant to us than ever. A perfect scheduling algorithm is the key to solve the load balance problems which can not only balance the load, but also can meet the users' needs. An optimal load balance algorithm is proposed in this paper. Algorithm proposed in this paper can enhance production of the systems and schedule the tasks to virtual machines (VMs) more efficiently. Finishing time of all tasks in the same system will be less than others'. The simulation tools is the CloudSim.

Keywords: Cloud Computing, Load balance, Interaction artificial bee colony, CloudSim.

1 Introduction

Cloud computing has recently emerged for its new hosting and delivering services over the Internet from "endpoint" to "cloud". It is a new style in which we need not to compute on the local computers, but on centralized facilities operated by third-party compute and storage utilities. At present, There is little consensus on the definition.

In this paper, we use one definition: A large-scale distributed computing paradigm driven by economies of scale, in which a pool of abstracted virtualized, dynamically-scalable, managed computing power, storage, platform and services are delivered on demand to external customers over the Internet[1].

In fact, Cloud is to outsource the provision of computing infrastructure to host services. Firstly, according to types of the service, there are three scenarios in the Cloud Computing model, including Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS). Furthermore, virtualization, parallel computing, grid computing and distributed computing are also important technologies to Clouds[2,3,4]. The basic idea of Cloud Computing can be shown as the following figure 1.

H. Sun et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 329, DOI: 10.1007/978-3-319-12286-1_6

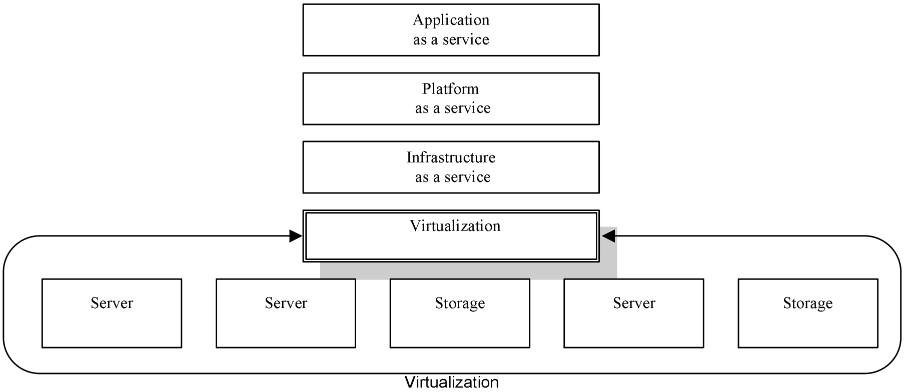


Fig. 1. The cloud computing scenarios diagram

Generally speaking, task scheduling algorithm of this paper is an optimal algorithm based on the interaction bee colony algorithm in cloud computing environment.

2 Related Work

For the cloud computing applications, the load balance in cloud computing has been widely used to improve the system response time and enhance utilization of system resources as possible. The first development period for Cloud Computing is called as the starting stage from 1996 to 1999, during which scholars found more opportunities. After that, the Cloud Computing has a development in enterprises until 2003. Furthermore, the period of the network upgrading emerged, starting in 2003 and ending 2005. The following three years, we encountered the load balancing upgrading and application development. In 2006, Google CEO Eric Schmidt first put forward the notion of Cloud Computing on SES San Jose 2006. After 2006, the load balance as an important part of the Cloud Computing has been widely spread.

How to design the algorithm is a key role to solve load balancing problems. The algorithm is related to the purpose of load balancing requirements and whether the system is good or not.

CloudSim, a Cloud Computing simulation software which was proposed by the University of Melbourne and Gridbus in 2007, by which we can establish the next generation computation center used to enable the flexible and dynamic platform for a virtual testing environment.

Generally speaking, there are three layers in cloud computing architecture on the CloudSim, including cloudlets layer, VMs layer and datacenter layer. The cloudlets layer is used to simulate the apps provided by users or service providers. The second layer, VMs layer, is virtual machines used to process tasks. The last one data center layer, is a Cloud Computing system. The following figure 2 is a Cloud Computing system diagram.

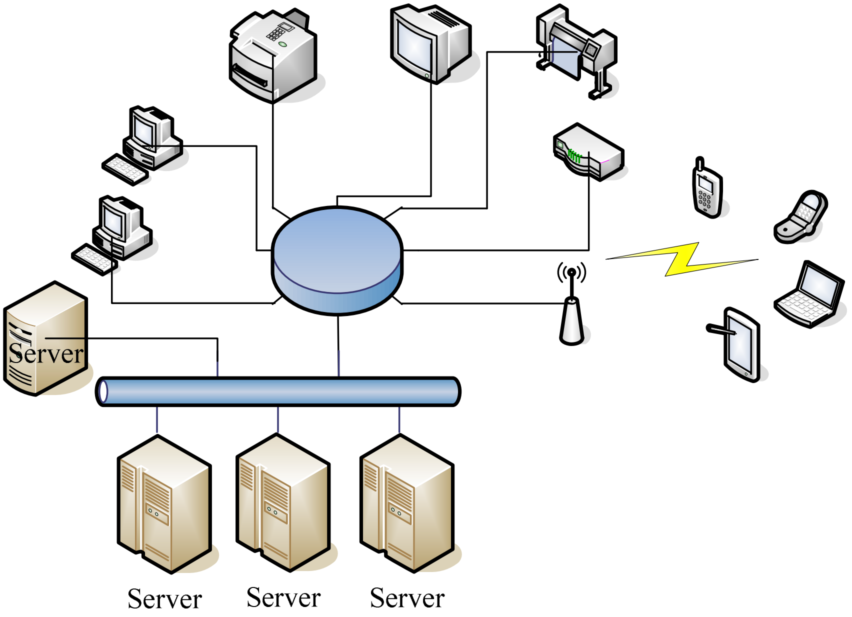


Fig. 2. The cloud computing system diagram

Actually, more and more intelligent algorithms are also introduced in the cloud computing research area. For example, M. Maheswaran mentioned dynamic heuristic allocation strategy in the task scheduling of distributed systems in the Hadoop [5]; Carretero solved task scheduling problem in the network by using the genetic algorithm [6], and got the minimum flow time and the shortest job completing time; jian Xiong put forward the dynamic model of a heterogeneous environment [7], it solves dynamic heterogeneous environment resource assignment problems through the interaction of swarm and environment; zem proposed an improved simulated annealing algorithm to solve the task scheduling problem in the network environment [8]; Chen Yulan put forward a constraint algorithm based on service quality, aiming to solve the of resource scheduling problems in grid computing [9]; Pandey S improved a kind of algorithm based on particle swarm optimization to solve the calculation and transmission system overhand in cloud computing [10]; Hua Xiayu [11] put forward an ant colony optimization scheduling algorithm for Cloud Computing.

3 Our Proposed Method

3.1 Environment Building

Firstly, we will create some cloudlets and VMs in order that the whole designed systems can be described. Furthermore, the parameters is set in advance in accordance to the actual system.

The identification of the VM is defined as vmid, and it is the unique symbol of the VM, which we can use it to distinguish different the VMs in the system. Furthermore, MIPS presents millions of instructions per second, it is used to show the computing speed, and is also a fundamental variable to measure the speed of CPUs. Moreover, the image size is defined as size. It is the storage size of a VM and we use it to set available space. In addition, the ram describes VM memory. While the computer is running, operation data is transferred to the memory. After the tasks are completed, the results will be sent from the ram. The BW is short for the bandwidth, which represents that how much data can be transmitted in a fixed period and capacity of transmitting data in the transport pipeline. Finally, the number of cups is pesnumber and the virtual machine manager name is VMm.

Just as the above, the parameters of the cloudlet can be described respectively. Firstly, the cloudlet is marked as id. It is the unique symbol of a cloudlet in the overall system, so that we can use it to distinguish different cloudlets. Secondly, the length of a cloudlet is recorded as length. It refers to the size of a cloudlet that if a VM can finish the applied task depending on. Furthermore, the size of a file is described as filesize. The size of an output file is outputsize. After tasks are completed, it figures the size of the cloudlets that will be output. Admittedly the utilization of a cloudlet is utilizationModel.

3.2 Mathematical Model

The number of cloudlets and the capacity of VMs express the load of virtual machines discussed in this paper [12]. As shown in the following formulation (1).

$$LOAD = MIPS * NumCPU + BW \tag{1}$$

As shown above, LOAD is the capacity of CPU, the millions of instructions per second is presented by MIPS, the NumCPU illustrates the number of cups, and the BW presents the bandwidth. The capacity of the load of a VM is mainly based on the MIPS, the number of CPU and the bandwidth.

The definition of load balance is shown below, which is based on the variance of the tasks lengths sum that is scheduled by every VM.

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

Here, P_i is the percentage of a VM tasks length in all VMs. F_i is the length of scheduled tasks in the VM numbered as i.

Then, a judgment formulation of load balance is given as shown in the following formulation (3).

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

Here, the vmNum is the number of all VMs and the variance represents average rate about the load of a VM accounting to the load of all VMs.

3.3 Tasks Scheduling Algorithm

For the scheduling algorithm is a NP problem, it means that algorithm cannot have access to the optimal solutions for all problems. However it can produce a wide range of optimal approximate solutions for the global optimal solutions [13].

I will describe the scheduling algorithm in this paper as follows:

1) Initialization: First of all, all parameters of VMs and cloudlets are set in accordance to the actual environment we need in our projects.

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

3) Judging balance: If the current system is under the balanced circumstance, the system directly distributes the task to the VM selected in the Step 2. Otherwise, go to the next step.

4) Iteration: After choosing a VM, due to the randomness of the VM selected in the Step 2, here, you can utilize the iteration formulation for searching a more efficient VM to implement the task. The iteration formulation, which we can find a more optimal solution according to, is shown in the following formulation (4).

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

As we can see, the iterative times is represented by t, and ϕ is a number between 0 and 1. The ϕ is calculated by the iteration formulation (5).

$$\phi = ((F_j * F_i)/(P_j - P_i))/\sum((F_m * F_n)/(P_m - P_n)) \tag{5}$$

In the formulation (5), we can acquire the parameter ϕin accordance to the gravity formulation developed by Newton, a famous British scientist. Through the iteration, a better solution to scheduling the task to VMs will be got than other solutions from a randomϕ. The formulation (5) is also the distinctive from the artificial bee colony algorithm in which the ϕ is a random number during 0 and 1. Because of that, we can get a more efficient and better solution. We can have a clear goal to search a better VM.

5) Judging overload: Through computing "bla", we can judge whether the current system is overloaded or not. According to the result, we decide whether we will use the limited formulation to balance the system or not. If the system is balanced, then go to the Step 7. Otherwise go to the Step 6.

6) Limited: If the result in the Step 4 presents the current system is overloaded, we will use the formulation below for renewing the system to schedule the task to another VM. The limited formulation is as show in the following formulation (6).

$$LOAD_i = LOAD_i + r * (LOAD_{\max} - LOAD_{\min}) \tag{6}$$

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

7) Judging balance: If the system is balanced, then judging the loop ends and distributing the tasks directly. Otherwise we make a decision to decide whether to go to Step 3 or not based on the result.

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

Generally speaking, an important advantage of the algorithm is that the system can be avoided to be overloaded, the algorithm makes all tasks scheduled more efficiently, and the tasks are scheduled to a better VM.

The whole process can be shown in the following figure 4.

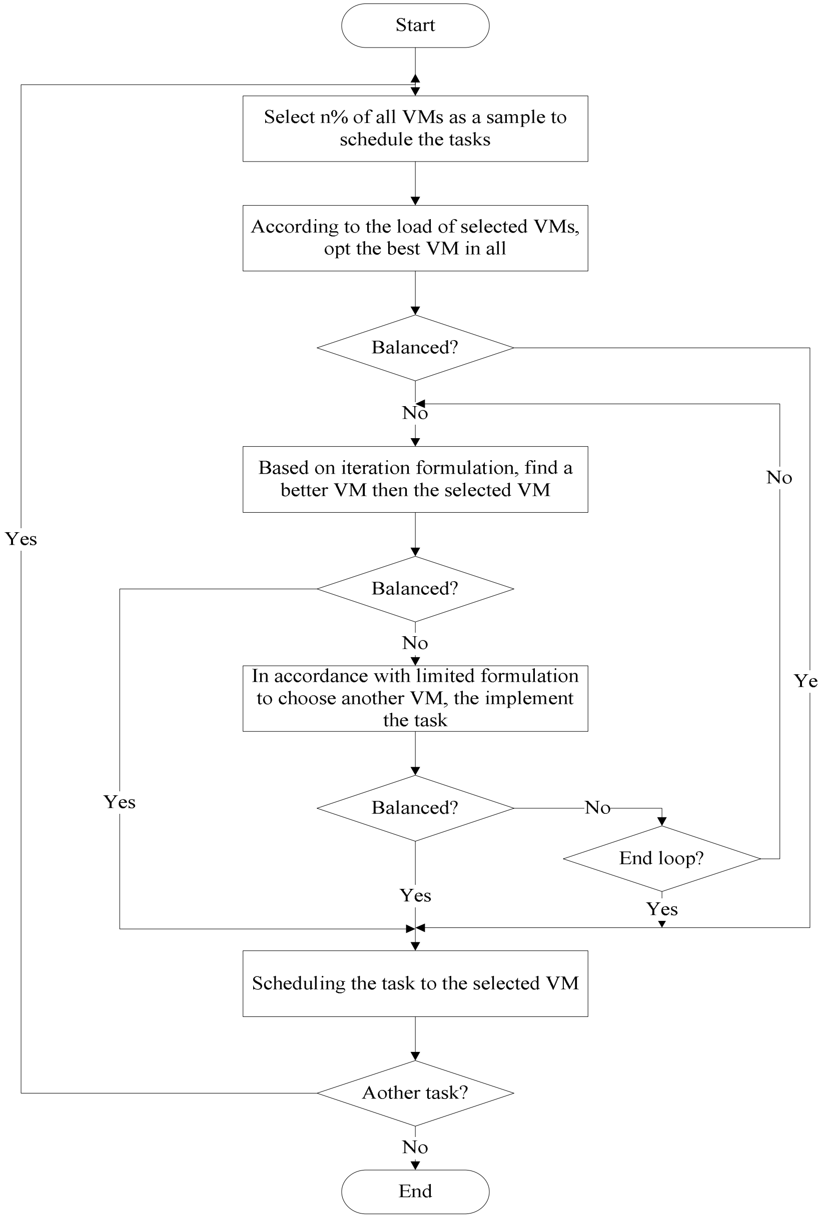

Start
Select n% of all VMs as a sample to schedule the tasks
According to the load of selected VMs, opt the best VM in all
Balanced?
No
Based on iteration formulation, find a better VM then the selected VM
Balanced?
No
In accordance with limited formulation to choose another VM, the implement the task
Balanced?
No
End loop?
No
Yes
Yes
Yes
Yes
Yes
Scheduling the task to the selected VM
Aother task?
No
End

Fig. 3. The flowchart of the ABC algorithm

In this paper, the balance criteria are to implement the program under the controlled circumstances, which can keep balance of the system in the current environment. At last, when all VMs are all overloaded and there are remaining cloudlets not to be processed, remnant tasks will be scheduled based on a random method.

4 Experimental Results and Analysis

In comparison to the ABC algorithm in which ϕ is a random number, not a digit calculated by interaction formulation, we can learn that the IABC algorithm is more efficient.

Firstly, one experiment gives the results when the number of cloudlets and the number of VMs change simultaneously. As shown in the following table 1.

Table 1. Parameters and results in the second experiment

VMs	10			20			30		
Cloudlets	100			200			300		
	Best	Worst	Average	Best	Worst	Average	Best	Worst	Average
IABC	1899	4525	3444	3721	7161	5188	5076	11364	6111
ABC	1957	9432	3632	4529	24110	11347	3808	37572	17918
VMs	40			50			60		
Cloudlets	400			500			600		
	Best	Worst	Average	Best	Worst	Average	Best	Worst	Average
IABC	5973	14585	9573	7506	16118	11416	9904	21932	12067
ABC	6929	50335	28429	6401	63061	21287	8836	51454	34977
VMs	70			80			90		
Cloudlets	700			800			900		
	Best	Worst	Average	Best	Worst	Average	Best	Worst	Average
IABC	18524	39672	25707	9836	14258	12393	11112	19706	13736
ABC	18710	33145	26458	10001	68567	20223	10409	23156	14524

Secondly, when VMs is a constant (60), the number of cloudlets increases. The results are in the above table 2.

Table 2. Parameters and results in the first experiment

Cloudlets	400			450			500		
	Best	Worst	Average	Best	Worst	Average	Best	Worst	Average
IABC	3891	8026	5101	5036	7660	6355	5611	10409	6775
ABC	3323	21855	5616	5205	29361	8781	5693	36868	9288
Cloudlets	550			600			650		
	Best	Worst	Average	Best	Worst	Average	Best	Worst	Average
IABC	6793	19839	10090	8830	12171	10843	9625	19842	14197
ABC	6695	64479	13177	7202	48433	13586	8735	58317	29550
Cloudlets	700			750			800		
	Best	Worst	Average	Best	Worst	Average	Best	Worst	Average
IABC	9754	20627	12405	11480	24571	16183	11082	21518	18652
ABC	8740	96049	44517	9766	93244	51490	12482	120102	54903

At last, the third experiment in which the number of cloudlets is constant (500), however the number of VMs is changing is introduced here. The results are in the following table 3.

Table 3. Parameters and results in the third experiment

VMs	20			30			40		
	Best	Worst	Average	Best	Worst	Average	Best	Worst	Average
IABC	11671	23331	14480	11018	20102	14134	7338	19219	13314
ABC	36630	83172	42570	19391	81276	37619	32084	70796	34598

VMs	50			60			70		
	Best	Worst	Average	Best	Worst	Average	Best	Worst	Average
IABC	7769	17687	11832	6562	10382	8102	6048	11008	7303
ABC	6154	62900	19134	7227	26080	11942	10156	23241	13974

VMs	80			90			100		
	Best	Worst	Average	Best	Worst	Average	Best	Worst	Average
IABC	3676	8346	6003	3305	6945	4787	2961	5819	3252
ABC	3894	16824	6412	2833	5524	4130	2441	5265	3151

As a result, from three above tables, we can learn that comparing with the Artificial Bee Colony (ABC) algorithm the IABC algorithm is more efficient in the best, worst and average result. Furthermore, the IABC algorithm also makes sure the system is under the balanced circumstance.

5 Conclusion

Based on the IABC algorithm, a load balance method to solve the load problems in Cloud Computing has been proposed. A large multitude of experimental results prove the efficiency of the algorithm. It is the first time that the IABC algorithm is introduced in solving such a load balance problems in Cloud Computing. Moreover, the CloudSim toolkit is used by us to simulate the actual system and test the algorithm. To modify the whole system design and finding a more efficient scheduling method is our future work.

Acknowledgment. The authors would like to thank for the support from the project NSFC (National Natural Science Foundation of China) with the Grant number 61202456, the Harbin Institute of Technology Innovation Fund project and the NDRC's cloud computing project.

References

1. Foster, I., Zhao, Y., Raicu, I., et al.: Cloud computing and grid computing 360-degree compared. In: Grid Computing Environments Workshop, GCE 2008, pp. 1–10. IEEE (2008)

2. Vaquero, L.M., Rodero-Merino, L., Caceres, J., et al.: A break in the clouds: towards a cloud definition. ACM SIGCOMM Computer Communication Review 39(1), 50–55 (2008)
3. Chang, B.R., Tsai, H.-F., Chen, C.-M.: Evaluation of Virtual Machine Performance and Virtualized Consolidation Ratio in Cloud Computing System. Journal of Information Hiding and Multimedia Signal Processing (JIHMSP) 4(3), 192–200 (2013)
4. Zhu, H., Liu, T., Zhu, D., Li, H.: Robust and Simple N-Party Entangled Authentication Cloud Storage Protocol Based on Secret Sharing Scheme. Journal of Information Hiding and Multimedia Signal Processing (JIHMSP) 4(2), 110–118 (2013)
5. Braun, T.D., Siegel, H.J., Beck, N., et al.: A comparison of eleven static heuristics for mapping a class of independent tasks onto heterogeneous distributed computing systems. The Journal of Parallel and Distributed Computing 61(6), 810–837 (2001)
6. Cañón, J., Alexandrino, P., Bessa, I., et al.: Genetic diversity measures of local European beef cattle breeds for conservation purposes. Genetics Selection Evolution 33(3), 311–332 (2001)
7. Jijian, L., Longjun, H., Haijun, W.: Prediction of vibration response of powerhouse structures based on LS-SVM optimized by PSO. Engineering Sciences 12, 009 (2011)
8. Kazem, A., Rahmani, A.M., Aghdam, H.H.: A modified simulated annealing algorithm for static task scheduling in grid computing. In: International Conference on Computer Science and Information Technology, ICCSIT 2008, pp. 623–627. IEEE (2008)
9. Yulan, J., Zuhua, J., Wenrui, H.: Multi-objective integrated optimization research on preventive maintenance planning and production scheduling for a single machine. International Journal of Advanced Manufacturing Technology 39(9-10), 954–964 (2008)
10. Pandey, S., Wu, L., Guru, S.M., et al.: A particle swarm optimization-based heuristic for scheduling workflow applications in cloud computing environments. In: 2010 24th IEEE International Conference on Advanced Information Networking and Applications (AINA), pp. 400–407. IEEE (2010)
11. Hua, X., Zheng, J., Hu, W.: Ant colony optimization algorithm for computing resource allocation based on cloud computing environment. Journal of East China Normal University (Natural Science) 1(1), 127–134 (2010)
12. Babu, L.D., Krishna, P.V.: Honey bee behavior inspired load balancing of tasks in cloud computing environments. Applied Soft Computing Journal 13(5), 2292–2303 (2013)
13. TSai, P.W., Pan, J.S., Liao, B.Y., et al.: Enhanced artificial bee colony optimization. The International Journal of Innovative Computing, Information and Control 5(12), 5081–5092 (2009)

Command Center Operation Miss Root Cause Analysis

XiaoLin Xiong and XinMing Xiang

CGN Power Co., Ltd.
Centre of Information Technology
ShenZhen, China
{xiongxiaolin,xiangxinming}@cgnpc.com.cn

Abstract. A command center is any place that is used to provide centralized command for some purpose. A command center enables an organization to function as designed, to perform day-to-day operations regardless of what is happening around it. There are parts of being human because no one acts perfectly correctly all the time. It's same with command center, we called it operation miss. We use RCA when it is released. Root Cause Analysis (RCA) is a problem-solving method used to identify and prove causes, design actions, and assess results. RCA is performed for pervasive issues that are identified through either persistent gaps or pervasive gaps.

Keywords: Command Center, Root Cause Analysis, Five Whys, fishbone diagram, Operation Miss.

1 Introduction

A command center is any place that is used to provide centralized command for some purpose. A command center enables an organization to function as designed, to perform day-to-day operations regardless of what is happening around it, in a manner in which no one realizes it is there but everyone knows who is in charge when there is trouble.

Conceptually, a command center is a source of leadership and guidance to ensure that service and order is maintained, rather than an information center or help desk. Its tasks are achieved by monitoring the environment and reacting to events, from the relatively harmless to a major crisis, using predefined procedures.

There are many types of command centers. In this article, we means date command center. A command center enables the real-time visibility and management of an entire service operation. Similar to an air traffic control center, a command center allows organizations to view the status of global service calls, service technicians, and service parts on a single screen. In addition, customer commitments or service level agreements (SLAs) that have been made can also be programmed into the command center and monitored to ensure all are met and customers are satisfied. The main job in command center is monitor and check server、network and other application system as schedule. Once we got a fatal、critical or warning alert, they deliver it to relevant supports to handle it. People who worked in command center are called as operators. All

H. Sun et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 329, DOI: 10.1007/978-3-319-12286-1_7

works in command center are carried out by job procedure instruction (JPI).We try our best to write the instructions prefect. Even then we always make misses in our daily work. We must analysis these misses or problems seriously. There are parts of being human because no one acts perfectly correctly all the time. It's same with command center, we called it operation miss. We use RCA when it is released.

2 Root Cause Analysis

Root Cause Analysis (RCA) is a problem-solving method used to identify and prove causes, design actions, and assess results. RCA is performed for pervasive issues that are identified through either persistent gaps or pervasive gaps. Most pervasive issues should be analyzed at the region or integrated marketing team level or above. However, all levels of the quality assurance organization should be familiar with the RCA method, which includes the following:

2.1 Understand the Problem

In order to be defined as a "problem," four pieces of information are required:

1. The actual current performance with historical trend detail (a picture is worth a thousand words!).
2. The desired performance (standard or goal).
3. The magnitude of the problem as seen by the difference between the actual and desired (sometimes referred to as the "gap").
4. The extent and characteristics of the problem or situation.

First, we must inspect the evidence. This requires a discipline to look at the process with a truly open mind and with no preconceptions.

Localizing and quantifying the impact. The first requirement of Problem Solving is to determine the merit of solving the problem.

Focusing on more than anecdotal evidence or individual situations

2.2 Describe Specific Possible Causes

A. We have to explore the chain of causes and effects and apply techniques like Whys?

Five Whys (why、why、why、why、why) is simple and easy to remember and also is good Rule of Thumb. Of course 5 is not always the number. We can take a dead car for an example,

My car will not start…

Why #1 ...why does my car not start? The battery is dead…

Why #2 ... why is the battery dead? The alternator is not working

Why #3 ... why is the alternator not working? The alternator belt was broken!

Why #4 ... why was the belt broken? It has never been replaced….

Why #5 … why was the belt never replaced? I was not maintaining my car according to the service schedule

We can got the root cause is human error, if we didn't analysis the root cause, all of us maybe just regards as the battery is the root cause.so we fix the battery, but we didn't eliminate hidden dangers.

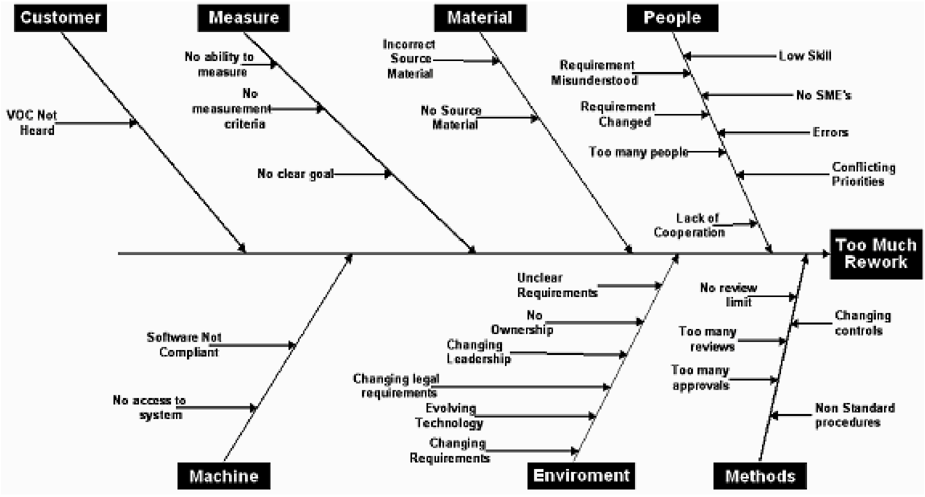
Complete a thorough root cause analysis- Five Whys

Example: The Washington Monument was disintegrating

1. Why is the Washington Monument disintegrating?
Use of harsh chemicals on a frequent basis
2. Why the harsh chemicals?
To clean pigeon droppings
3. Why so many pigeons?
They eat spiders and there are a lot of spiders at monument
4. Why so many spiders?
They eat gnats and lots of gnats at monument
5. Why so many gnats?
They are attracted to the light at dusk.
Solution: Turn on the lights at a later time, means after dusk.
B. Forming specific hypotheses or cause descriptions.

A technique for identifying the root cause of a problem is to use an Ishikawa diagram, also referred to as a cause-and-effect diagram or fishbone diagram. An Ishikawa diagram is typically the result of a brainstorming session in which members of a group offer ideas to improve a product. For problem-solving, the goal will be to find the cause(s) of the problem.

First there is the main subject, which is the backbone of the diagram that we are trying to solve or improve. The main subject is derived from a cause. The relationship between a cause and an effect is a double relation: an effect is a result of a cause, and the cause is the root of an effect.



2.3 Prove and Act on the Real Causes

Getting the data that proves or disproves hypotheses

For proven causes, designing specific actions

Assigning an owner and tracking progress

Team Review is a Critical Part of the Analysis Phase. Recall the question that opened this section:

Why Did You Pick Up This Problem? In the review, the problem solver, the managers, and the team meet to ensure that all aspects of the problem have been considered.

2.4 Assess Results and Analyze Feedback

The desired behaviors when applying the RCA method to the SSL cadence include:

Use in-cadence reviews of similar transaction data across all levels versus using ad hoc reviews

Use tools, data, and upward information flow to eliminate ad hoc reviews and top-down reviews

Coach individuals to leverage their strengths and improve results during cadence reviews.

Coach individuals to be more productive, accountable, and effective

Focus on building trust and enhancing team dynamics during the cadence (leader behavior)

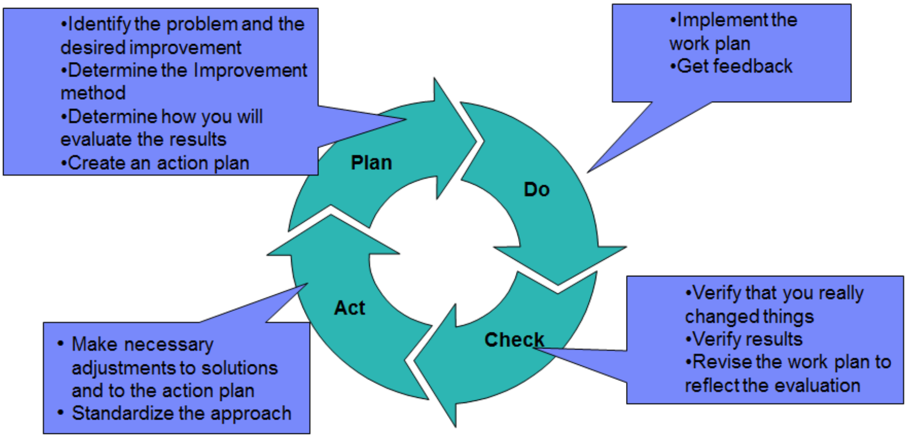
Coach on pervasive issues and key transactions

Schedule and execute cross-team integration meetings (client team behavior)

3 Consider Alternative Solutions While Building Consensus

There is always more than one way to solve any problem and places high value on the creativity of the problem solver in their culture, therefore, There aren’t a lot of absolutes around this part of the process.

Here are some guidelines for the evaluation process. Broadly consider all possibilities. Narrow the list by eliminating impractical solutions or combining similar items. Evaluate based on simplicity, cost, area of control, and the ability to implement quickly. Develop consensus on the proposed solution. Test ideas for effectiveness. Select the best solution. Plan-Do-Check-Act


•Identify the problem and the desired improvement
•Determine the Improvement method
•Determine how you will evaluate the results
•Create an action plan
Plan
•Implement the work plan
•Get feedback
Do
Act
Check
•Verify that you really changed things
•Verify results
•Revise the work plan to reflect the evaluation
• Make necessary adjustments to solutions and to the action plan
• Standardize the approach

The evaluation criteria force teams to find feasible solutions quickly. Is it within your control to implement? Can you do it without outside support? Is it possible to implement the solution quickly? (Today is best.) Is the solution a simple and effective one? Is the solution low cost, or even better, no cost?

There is a tendency toward “fancy” or “high-tech” solutions to problems. Invariably the latest technology or machine is suggested. In rare cases the technology is needed; however, while waiting for the “ultimate” solution, consider a short-term improvement that can be implemented immediately.

4 Develop Consensus and Test Ideas for Effectiveness

Because the choices have been thoroughly analyzed, it is usually easy to develop consensus, but if not, proceed to the next step and test each idea to find the best one. Consensus doesn’t mean complete agreement. rather, it means that everyone agrees to accept the proposed solution.

Testing ideas can come in many forms, but it’s important to do validation like this before implementing a proposed solution. Paper simulations, mock-ups, etc. can all serve to help testing.

Always consider short-term temporary countermeasures for immediate benefits. Divide larger tasks into smaller segments, with assigned completion dates and measurements for each portion.

Responsibility for an action item does not mean that the responsible person has to do the task. They are responsible for the outcome and for ensuring progress. The only way to verify results is to ensure that an effective measurement process is in place prior to implementation so that a before and after comparison can be made.

Once your solutions become a reality, it will probably be necessary to make adjustments. Carefully observe the new process to verify that it is free from major problems. Always conclude your process with a look to the future. Continual improvement means forever! Set the expectation that the process of improvement is never complete.

What are pervasive issues? Pervasive issues are problems that are common across the process, not problems due to special or isolated issues. Pervasive issues are identified through either pervasive gaps or persistent gaps. Pervasive gaps are identified as a result of the roll-up of information from the SSL cadence. For example, an issue might arise in the cadence that has impact across the majority of brands or sectors in a single geography or region. Another example is an issue that impacts a single business in multiple regions. The issues that are constraints or barriers to effective execution by the sales teams must also be identified; these constraints and barriers are often pervasive issues.

Persistent gaps become evident during Exception Management and reports from meetings, such as the Global Sales Integration Review (GSIR) or Geo/IOT Sales Reviews. For example, Exception Management identifies Business Units, Geos, and brands that consistently miss their targets for at least a quarter.

What is the desired behavior when an issue is identified in the cadence? Instead of having an immediate ask questions to coach the desired behavior. Is this issue pervasive? If yes, why did it occur?

Once you understand why, take the necessary actions and assess results.

The following questions can help determine if an issue is pervasive and if it needs analysis and resolution.

Was this a specific incident (isolated) or a common problem (pervasive)?

When did it occur? How frequently?

Where did it occur (physical location)?

Is this issue constraining the field? How much will sales execution improve by removing this barrier?

How significant is this problem (revenue, client satisfaction, or employee morale

impact)?

Does the problem warrant the resources needed to study it?

5 Conclusion

The benefits of using the RCA method include: Managing by exception - less ad-hoc inspection.

Experiencing better collaboration when making business decisions - RCA directly supports Sales Leadership competencies (collaborative influence, earning trust, and informed judgment) when properly applied

Improving coaching

Focusing on execution; enabling individuals to be more productive, accountable, and effective during the cadence.

References

[1] Okes, D.: Root Cause Analysis: The Core of Problem Solving and Corrective Action Hardcover (2009)

[2] Andersen, B.: The ASQ Pocket Guide to Root Cause Analysis Spiral-bound (2013)

[3] Latino, R.J.: Root Cause Analysis: Improving Performance for Bottom-Line Results, 4th edn. (2011)

Part II

Recent Advances on Evolutionary Optimization Technologies

A New Cat Swarm Optimization with Adaptive Parameter Control

Jianguo Wang

Teaching Affairs Office, Taishan University, Taian 271021, P.R. China
tsuwjg@163.com

Abstract. Cat Swarm Optimization (CSO) is a new swarm intelligence based algorithm, which simulates the behaviors of cats. In CSO, there are two search modes including seeking and tracing. For each cat (solution) in the swarm, its search mode is determined by a parameter *MR* (mixture ratio). In this paper, we propose a new CSO algorithm by dynamically adjusting the parameter *MR*. In addition, a Cauchy mutation operator is utilized to enhance the global search ability. To verify the performance of the new approach, a set of twelve benchmark functions are tested. Experimental results show that the new algorithm performs better than the original CSO algorithm.

Keywords: Cat swarm optimization (CSO), swarm intelligence, adaptive parameter, Cauchy mutation, Global optimization.

1 Introduction

In real world, many application problems can be converted into optimization problems over continuous or discrete search space. To efficiently solve optimization problems, some swarm intelligence based algorithms have been proposed, such as Particle Swarm Optimization (PSO) [1], Ant Colony Optimization (ACO) [2], Artificial Bee Colony (ABC) [3], and Cat Swarm Optimization (CSO) [4], etc.

CSO was firstly proposed by Chu and Tsai [4], which is inspired by the behaviors of cats. In the past several years, CSO has been applied to various optimization fields. Santosa and Ningrum [5] applied CSO to for clustering. Simulation results indicate that the CSO outperforms K-means and PSO in terms the accuracy of clustering. Kumar and Kalavathi [6] presented an application of CSO on the optimal placement of multiple UPFC's in voltage stability enhancement under contingency. Panda et al. [7] used CSO for IIR system identification. Results demonstrate superior identification performance of CSO compared to that achieved by GA and PSO. In [8], CSO was applied to determine the best optimal impulse response coefficients of FIR low pass, high pass, band pass and band stop filters, trying to meet the respective ideal frequency response characteristics. The results show that CSO achieves better performance when compared with real coded GA (RGA), PSO, and differential evolution (DE). Pardhan and Panda [9] used CSO to solve multi-objective optimization problems. Simulation results demonstrate that the CSO can be a better

H. Sun et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 329, DOI: 10.1007/978-3-319-12286-1_8

candidate for solving multi-objective problems. In [10], Tsai et al. proposed a parallel CSO with information exchanging (PCSO), in which the virtual cats share the isolated near best solution between different clusters via the information exchanging process.

In this paper, we propose a new CSO algorithm called NCSO, which employs a dynamical method to adjust the parameter *MR* during the evolution. Experiments are conducted on twelve benchmark functions. Simulation results show that the NCSO outperforms the original CSO on the majority of test functions.

The rest paper is organized as follows. In Section 2, the original CSO algorithm is briefly introduced. In Section 3, the parameter *MR* is investigated. In Section 4, our algorithm is proposed. Simulation results are presented in Section 5. Finally, the work is concluded in Section 6.

2 Cat Swarm Optimization

Each cat in the swarm can be considered as a candidate solution. There are two search modes, seeking and tracing, for each cat. During the search process, cats are randomly selected from the swarm and their flags are set to seeking mode or tracing mode. This selection operation is based on a parameter *MR*. In CSO, these two modes of operations are mathematically modeled for solving complex optimization problems [7].

2.1 Seeking Mode

The seeking mode is used to model the cat during a period of resting but being alert – looking around its environment for its next move [4]. It mainly contains the following basic factors: seeking memory pool (*SMP*), seeking range of selected dimension (*SRD*), counts of dimension to change (*CDC*), and self position consideration (*SPC*). The *SMP* indicates the number of copies of a cat produced in the seeking mode. The *SRD* declares the mutative ration for the selected dimensions. While in seeking mode, if a dimension is selected for mutation, the difference between the old and new values may not be out of range. The *CDC* is the number of dimensions to be changed. The SPC is a Boolean valued variable, which represents whether the point at which the cat is already standing will be one of the candidate points to move to.

The seeking mode can be described as follows [7].

Step 1. For the *i*th solution (cat), make *SMP* copies.

Step 2. Based on the parameter *CDC*, update the position of each copy by randomly adding or subtracting *SRD* percents the current position value.

Step 3. Compute the fitness values of all copies.

Step 4. Select the best candidate from the *SMP* copies as the position of the *i*th cat.

2.2 Tracing Mode

The tracing mode corresponds to a local search technique for the optimization problem [7]. In this mode, a cat chases the target according to its own velocity. The tracing operation is similar to the movement of particles in PSO. In tracing model, each cat a position vector (*X*) and a velocity vector (*V*). During the search process, a cat moves according to the following equations [4].

$$V_{id} = w \cdot V_{id} + c \cdot r \cdot \left(P_{gd} - X_{id} \right). \tag{1}$$

$$X_{id} = X_{id} + V_{id}. \tag{2}$$

where V_{id} is the velocity of the *i*th cat on the *d*th dimension, X_{id} is the position of the *i*th cat on the *d*th dimension, P_{gd} is the *d*th dimension of the global best position. The parameter *w* is the inertia weight, *c* is the acceleration constant, and *r* is a random number with the range of [0, 1].

2.3 Framework of CSO

The main steps of CSO are described as below [7].

Step 1: Randomly generate *N* cats (solutions) in the swarm, and calculate their fitness values. This operation consists of the initialization of positions and velocities of all cats.
Step 2. Select the best cat (with the best fitness value) as the P_g.
Step 3. According to the mixture ration (*MR*), cats are randomly selected from the swarm, and their flags are set to seeking mode or tracing mode.
Step 4. For each cat, if its flag is seeking mode, conduct the seeking mode process (according to Section 2.1); otherwise execute the tracing mode operation (according to Section 2.2).
Step 5. Calculate the fitness values of all cats.
Step 6. Update P_g, if possible.
Step 7. If the stopping condition is satisfied, terminate the algorithm and output the results; otherwise go to Step 3.

3 Parameter Analysis of *MR*

As mentioned before, there are two search modes for CSO. How to choose the search mode is determined by the mixture ratio (*MR*). To investigate the effects of *MR* on the performance of CSO, this section presents a parameter analysis of *MR*. In the experiment, the parameter *MR* is set to 0, 0.1, 0.2, 0.3, 0.5, and 1.0, respectively. The computational results of CSO under different values of *MR* are compared. For other parameters of CSO, we use the following settings. The population size and the maximum number of fitness evaluations are set to 30 and 1.0e+05, respectively. The *SRD*, *CDC*, *w*, and *c* are set to 20%, 90%, 0.73, and 2.0, respectively. For each test function, CSO is run 30 times, and the mean best fitness value is reported.

3.1 Test Functions

There are twelve well-known benchmark functions used in the following experiments. These problems were utilized in previous studies [11]. According to their properties, they are divided into three groups: unimodal and simple multimodal functions (f_1-f_2), unrotated multimodal functions (f_3-f_8), and rotated multimodal functions (f_9-f_{12}). All test functions are minimization problems. In this paper, we only consider the problems with D=30. For the rotated problems, the original variable x is left multiplied by the orthogonal matrix **M** to get the new rotated variable y=**M***x. The mathematical descriptions of these functions are described as follows.

1) Sphere function

$$f_1(x)=\sum_{i=1}^{D} x_i^2$$

where $x_i \in [-100, 100]$, and the global optimum is 0.

2) Rosenbrock's function

$$f_2(x)=\sum_{i=1}^{D-1}\left[100\left(x_i^2-x_{i+1}\right)^2+\left(x_i-1\right)^2\right]$$

where $x_i \in [-2.048, 2.048]$, and the global optimum is 0.

3) Ackley's funcction

$$f_3(x)=-20\exp\left(-0.2\sqrt{\frac{1}{D}\sum_{i=1}^{D} x_i^2}\right)-\exp\left(\frac{1}{D}\sum_{i=1}^{D}\cos\left(2\pi x_i\right)\right)+20+e$$

where $x_i \in [-32.768,32.768]$, and the global optimum is 0.

4) Griewanks's function

$$f_4(x)=\sum_{i=1}^{D}\frac{x_i^2}{4000}-\prod_{i=1}^{D}\cos\left(\frac{x_i}{\sqrt{i}}\right)+1$$

where $x_i \in [-600,600]$, and the global optimum is 0.

5) Weierstrass function

$$f_5(x)=\sum_{i=1}^{D}\left(\sum_{k=0}^{k\max}\left[a^k\cos\left(2\pi b^k\left(x_i+0.5\right)\right)\right]\right)$$
$$-D\sum_{k=0}^{k\max}\left[a^k\cos\left(2\pi b^k\cdot 0.5\right)\right]$$
$$a=0.5, b=3, k\max=20$$

where $x_i \in [-0.5,0.5]$, and the global optimum is 0.

6) Rastrigin's function

$$f_6(x)=\sum_{i=1}^{D}\left(x_i^2-10\cos\left(2\pi x_i\right)+10\right)$$

where $x_i \in [-5.12,5.12]$, and the global optimum is 0.

7) Noncontinuous Rastrigin's function

$$f_7(x)=\sum_{i=1}^{D}\left(y_i^2-10\cos\left(2\pi y_i\right)+10\right)$$
$$y_i=\begin{cases} x_i, & \left|x_i\right|<\frac{1}{2} \\ \frac{round\left(2x_i\right)}{2}, & \left|x_i\right|\geq\frac{1}{2}\end{cases}$$

where $x_i \in [-5.12,5.12]$, and the global optimum is 0.

8) Schwefel's function

$$f_8\left(x\right)=418.9829\cdot D-\sum_{i=1}^{D}\left(x_i\sin\left(\sqrt{\left|x_i\right|}\right)\right)$$

where $x_i \in [-500,500]$, and the global optimum is 0.
9) Rotated Ackley's function

$$f_9\left(x\right)=-20\exp\left(-0.2\sqrt{\frac{1}{D}\sum_{i=1}^{D}y_i^2}\right)-\exp\left(\frac{1}{D}\sum_{i=1}^{D}\cos\left(2\pi y_i\right)\right)+20+e$$

where $x_i \in [-32.768,32.768]$, and the global optimum is 0.
10) Rotated Griewanks's function

$$f_{10}\left(x\right)=\sum_{i=1}^{D}\frac{y_i^2}{4000}-\prod_{i=1}^{D}\cos\left(\frac{y_i}{\sqrt{i}}\right)+1$$

where $x_i \in [-600,600]$, and the global optimum is 0.
11) Rotated Weierstrass function

$$f_{11}\left(x\right)=\sum_{i=1}^{D}\left(\sum_{k=0}^{k\max}\left[a^k\cos\left(2\pi b^k\left(y_i+0.5\right)\right)\right]\right) \\ -D\sum_{k=0}^{k\max}\left[a^k\cos\left(2\pi b^k\cdot 0.5\right)\right]$$

where $x_i \in [-0.5,0.5]$, and the global optimum is 0.
12) Rotated Rastrigin's function

$$f_{12}\left(x\right)=\sum_{i=1}^{D}\left(y_i^2-10\cos\left(2\pi y_i\right)+10\right)$$

where $x_i \in [-5.12,5.12]$, and the global optimum is 0.

Table 1. The mean fitness values of CSO under different *MR*

Functions	*MR*=0.0	*MR*=0.1	*MR*=0.2	*MR*=0.3	*MR*=0.5	*MR*=1.0
	Mean	Mean	Mean	Mean	Mean	Mean
f_1	1.06E-16	**1.16E-24**	3.00E-20	1.83E-16	2.13E-13	1.35E-23
f_2	**2.10E+01**	2.25E+01	2.39E+01	2.46E+01	2.58E+01	2.75E+01
f_3	3.16E+00	**2.24E-13**	5.77E-11	4.91E-09	2.78E-07	9.90E-12
f_4	9.99E-16	1.97E-02	1.72E-02	2.70E-02	1.97E-02	**0.00E+00**
f_5	3.37E+00	**1.42E-14**	5.78E-09	2.10E-04	2.79E-03	2.96E-02
f_6	4.88E+01	3.88E+01	4.18E+01	**2.09E+01**	2.21E+01	1.61E+02
f_7	**1.42E+01**	2.10E+01	5.83E+01	2.43E+01	1.73E+01	2.23E+02
f_8	4.84E+03	**4.34E+03**	5.09E+03	6.98E+03	5.63E+03	8.43E+03
f_9	3.09E+00	**1.25E-12**	1.32E-10	5.62E-09	7.43E-07	3.36E-12
f_{10}	1.97E-02	6.58E-02	**0.00E+00**	2.21E-02	3.21E-02	**0.00E+00**
f_{11}	1.43E+01	8.43E+00	**5.43E-05**	4.61E-04	1.06E-02	5.25E-01
f_{12}	5.64E+01	8.56E+01	6.77E+01	**2.69E+01**	4.39E+01	2.20E+02

3.2 Results for CSO with Different *MR*

Table 1 presents the computational results of CSO under different values of *MR*, where "Mean" indicates the mean fitness value. The best results are shown in bold. It can be seen that *MR*=0.1 and *MR*=1.0 are suitable for function f_1. For any *MR*, CSO can not find reasonable solutions on f_2, f_6, f_7, f_8, and f_{12}. For function f_3, CSO with *MR*=0.0 falls into local minima, while other cases can find near-optimal solutions. There are two extreme cases, *MR*=0.0 and *MR*=1.0. For *MR*=0.0, there is only tracing mode is CSO. All cats trace the target during the search process. For *MR*=1.0, all cats follow the seeking mode to search candidate solutions. It is obvious that CSO with a single tracing or seeking mode obtains poor performance. However, the single search mode may be suitable for some specific problems, such as f_2, f_4, f_9, and f_{10}. We can not find a fixed value of *MR* that suitable for solving all test functions. It means that the parameter *MR* is problem-oriented. Results show that 0<*MR*<0.3 may be a good choice the test suite.

4 Proposed Approach

In Section 3, we present an experimental study on the parameter analysis of *MR*. The results show that there is no fixed value of *MR* for all test functions. For different test problems, different values of *MR* are required. *MR* between 0 and 0.3 may help CSO achieve good results. It seems that CSO with dynamical *MR* may be a good choice.

Based on the above analysis, we propose a simple method to dynamically adjust the *MR* during the search process. At the beginning of each iteration, the parameter *MR* is updated as follows.

$$MR = rand(0, 0.3) \,. \tag{3}$$

where *rand*(0, 0.3) is a random value between 0 and 0.3.

To enhance the global search of CSO, a Cauchy mutation operator is employed. The Cauchy jump is not a new technique, and it has been applied to other optimization algorithms [11]. The main idea behind Cauchy jump is conducting a Cauchy mutation on the global best firefly. It is hopeful that the long fat tails of Cauchy distribution can help trapped fireflies jump to better positions [11].

The one-dimensional Cauchy density function centered at the origin is defined by [11]:

$$f(x) = \frac{1}{\pi}\frac{t}{t^2 + x^2} \,. \tag{4}$$

where t>0 is a scale parameter. The Cauchy distribution function is [11]:

$$F(x) = \frac{1}{2} + \frac{1}{\pi}\arctan\left(\frac{x}{t}\right) . \tag{5}$$

The Cauchy mutation used in this paper is defined by

$$P^*_{gd} = P_{gd} + cauchy() \cdot \tag{6}$$

where P_{gd} is the dth dimension of the global best solution, and *cauchy*() is a random value generated by the Cauchy distribution function with the scale parameter t=1.

The main steps of our new approach NCSO are listed as follows.

Step 1: Randomly generate N cats (solutions) in the swarm, and calculate their fitness values. This operation consists of the initialization of positions and velocities of all cats.

Step 2. Select the best cat (with the best fitness value) as the P_g.

Step 3. Update the parameter *MR* according to Eq. (3).

Step 4. According to the mixture ration (*MR*), cats are randomly selected from the swarm, and their flags are set to seeking mode or tracing mode.

Step 5. For each cat, if its flag is seeking mode, conduct the seeking mode process (according to Section 2.1); otherwise execute the tracing mode operation (according to Section 2.2).

Step 6. Calculate the fitness values of all cats.

Step 7. Update P_g, if possible.

Step 8. Conduct the Cauchy mutation according to Eq. (6). If the P_g^* is better than P_g, replace P_g with P_g^*.

Step 9. If the stopping condition is satisfied, terminate the algorithm and output the results; otherwise go to Step 3.

Table 2. Comparison results between CSO and NCSO

	CSO						NCSO
Functions	*MR*=0.0	*MR*=0.1	*MR*=0.2	*MR*=0.3	*MR*=0.5	*MR*=1.0	
	Mean	Mean	Mean	Mean	Mean	Mean	Mean
f_1	1.06E-16	**1.16E-24**	3.00E-20	1.83E-16	2.13E-13	1.35E-23	1.68E-21
f_2	**2.10E+01**	2.25E+01	2.39E+01	2.46E+01	2.58E+01	2.75E+01	2.35E+01
f_3	3.16E+00	**2.24E-13**	5.77E-11	4.91E-09	2.78E-07	9.90E-12	6.54E-12
f_4	9.99E-16	1.97E-02	1.72E-02	2.70E-02	1.97E-02	**0.00E+00**	**0.00E+00**
f_5	3.37E+00	**1.42E-14**	5.78E-09	2.10E-04	2.79E-03	2.96E-02	1.73E-10
f_6	4.88E+01	3.88E+01	4.18E+01	**2.09E+01**	2.21E+01	1.61E+02	7.66E+01
f_7	1.42E+01	2.10E+01	5.83E+01	2.43E+01	1.73E+01	2.23E+02	**6.38E-04**
f_8	4.84E+03	4.34E+03	5.09E+03	6.98E+03	5.63E+03	8.43E+03	**3.56E+03**
f_9	3.09E+00	**1.25E-12**	1.32E-10	5.62E-09	7.43E-07	3.36E-12	2.08E-11
f_{10}	1.97E-02	6.58E-02	**0.00E+00**	2.21E-02	3.21E-02	**0.00E+00**	3.70E-02
f_{11}	1.43E+01	8.43E+00	**5.43E-05**	4.61E-04	1.06E-02	5.25E-01	5.74E+00
f_{12}	5.64E+01	8.56E+01	6.77E+01	**2.69E+01**	4.39E+01	2.20E+02	5.57E+01
w/t/l	**9/0/3**	**6/0/6**	**9/0/3**	**8/0/4**	**8/0/4**	**7/1/4**	

5 Simulation Results

In this section, the performance of our new CSO (NCSO) algorithm is compared with the original CSO under different values of *MR*. To have a fair comparison, the same parameter settings are used. For both CSO and NCSO, the population size and the maximum number of fitness evaluations are set to 30 and 1.0e+05, respectively. The *SRD*, *CDC*, *w*, and *c* are set to 20%, 90%, 0.73, and 2.0, respectively. For CSO, the parameter *MR* is set to different values as described in Section 3. For NCSO, the *MR* is adaptive, and we do not need to manually set it. For each test function, each algorithm is conducted 30 runs, and the mean best fitness value is reported.

Table 2 gives the computational results of NCSO and CSO with different *MR*, where "*w*/*t*/*l*" means that NCSO wins in *w* functions, ties in *t* functions, and loses in *l* functions. As shown, NCSO outperforms CSO with *MR*=0.1 and *MR*=0.2 on 9 functions, while NCSO achieves worse results on 3 functions. Both CSO with *MR*=0.1 and NCSO win 6 functions. It seems that they achieve similar performance. For *MR*=0.3 and 0.5, NCSO performs better than CSO on 8 functions. For the rest of 4 functions, CSO is better than NCSO. When *MR*=1.0, NCSO achieves better results than CSO on 7 functions, while CSO outperforms NCSO on 4 functions. Both of them can converge to the global optimum on f_4. Fig. 1 shows the convergence characteristics of NCSO with other six CSO algorithms on two representative functions. It can be seen that NCSPO converges faster than other algorithms.

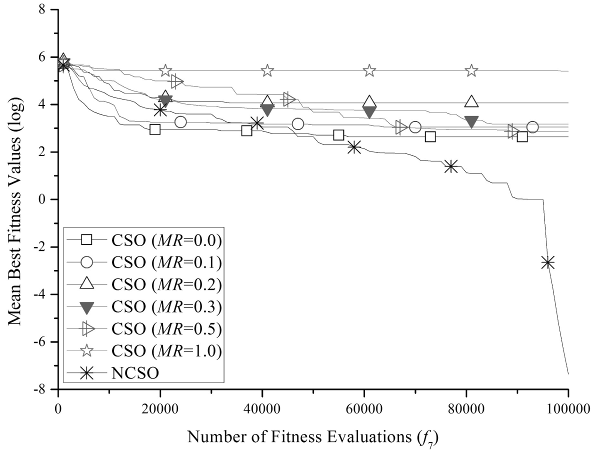
To clearly compare the performance of NCSO and other six versions of CSO, Friedman test is used to calculate the average rankings of these algorithms on the test suite [12]. Table 3 presents the results of the average rankings. The best ranking (with the lowest ranking value) is shown in bold. Results show that NCSO achieves the best ranking and CSO with *MR*=0.1 obtains the second place. It demonstrates that NCPSO performs better than six versions of CSO. That also confirms the efficiency of the proposed method.

Table 3. Average rankings achieved by Friedman test

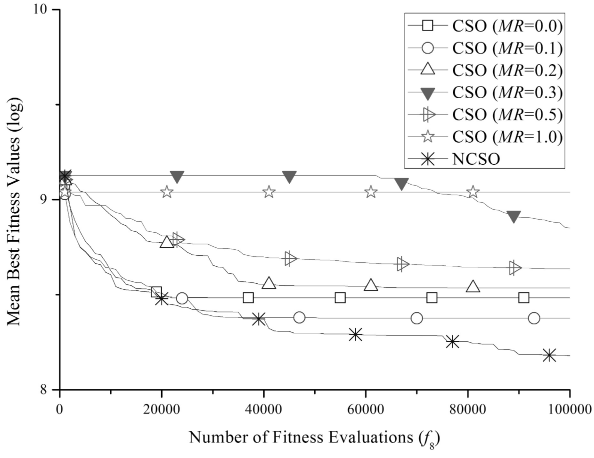
Algorithms	Average Rankings
NCSO	**3.04**
CSO (*MR*=0.0)	4.50
CSO (*MR*=0.1)	3.29
CSO (*MR*=0.2)	3.71
CSO (*MR*=0.3)	4.25
CSO (*MR*=0.5)	4.63
CSO (*MR*=1.0)	4.58

6 Conclusion

The mixture ratio (*MR*) is an important parameter which controls the seeking and tracing modes. To reduce the effects of *MR* on the performance of CSO, this paper proposes a new CSO algorithm (NCSO) by employing a dynamical method to adjust the parameter *MR*. Moreover, the Cauchy mutation operator is utilized to enhance the global search. Experiments are conducted on twelve test functions. Simulation results show that the proposed parameter method can effectively improve the performance of CSO.



Noncontinuous Rastrigin's function (f_7)



Schwefel's function (f_8)

Fig. 1. The convergence characteristics of NCSO with CSO on two representative functions

References

1. Kennedy, J., Eberhart, R.C.: Particle Swarm Optimization. In: Proceedings of International Conference on Neural Networks, vol. IV, pp. 942–1948. IEEE Press, Piscataway (1995)
2. Dorigo, M., Maniezzo, V., Colorni, A.: The Ant System: Optimization by a Colony of Cooperating Agents. IEEE Transactions on Systems, Man and Cybernetics-Part B: Cybernetics 26(1), 29–41 (1996)
3. Karaboga, D.: An Idea Based on Honey Bee Swarm for Numerical Optimization. Technical Report-TR06, Erciyes University, Engineering Faculty, Computer engineering Department (2005)
4. Chu, S.C., Tsai, P.W.: Computational Intelligence Based on the Behavior of Cats. International Journal of Innovative Computing, Information and Control 3, 163–173 (2007)
5. Santosa, B., Ningrum, M.K.: Cat Swarm Optimization for Clustering. In: Proceedings of International Conference of Soft Computing and Pattern Recognition, pp. 54–59 (2009)
6. Kumar, G.N., Kalavathi, M.S.: Cat Swarm Optimization for Optimal Placement of Multiple UPFC's in Voltage Stability Enhancement under Contingency. International Journal of Electrical Power and Energy Systems 57, 97–104 (2014)
7. Panda, G., Pradhan, P.M., Majhi, B.: IIR System Identification Using Cat Swarm Optimization. Expert Systems with Applications 38, 12671–12683 (2011)
8. Saha, S.K., Ghoshal, S.P., Kar, R., Mandal, D.: Cat Swarm Optimization Algorithm for Optimal Linear Phase FIR Filter Design. ISA Transactions 52, 781–794 (2013)
9. Pradhan, P.M., Panda, G.: Solving Multiobjective Problems Using Cat Swarm Optimization. Expert Systems with Applications 39, 2956–2964 (2012)
10. Tsai, P.W., Pan, J.S., Chen, S.M., Liao, B.Y., Hao, S.P.: Parallel Cat Swarm Optimization. In: Proceedings of International Conference on Machine Learning and Cybernetics, pp. 6309–6319 (2008)
11. Wang, H., Wu, Z.J., Rahnamayan, S., Liu, Y., Ventresca, M.: Enhancing Particle Swarm Optimization Using Generalized Opposition-Based Learning. Information Sciences 181(20), 4699–4714 (2011)
12. Wang, H., Rahnamayan, S., Sun, H., Omran, M.G.H.: Gaussian Bare-Bones Differential Evolution. IEEE Transactions on Cybernetics 43(2), 634–647 (2013)

Multi-objective Nondominated Sorting Invasive Weed Optimization Algorithm for the Permanent Magnet Brushless Direct Current Motor Design

Si-Ling Wang[1,*], Bao-Wei Song[1], and Gui-Lin Duan[2]

[1] College of Marine Engineering, Northwestern Polytechnical University,
Xi'an 710072, China
276344371@qq.com
[2] CSIC710, Yichang Hubei 443003, P.R. China

Abstract. In this paper, we proposed a new multi-objective optimization algorithm named Nondominated Sorting Invasive Weed Optimization (NSIWO) which was inspired from Nondominated Sorting Genetic Algorithm II(NSGA-II) and Invasive Weed Optimization (IWO). Firstly, the fast nondominated sorting algorithm was used to rank the weeds, and the number of seeds produced by a weed increased linearly from highest rank to the lowest rank. Moreover, in order to get a good distribution and spread of Pareto-front, crowding distance was used for determining the seeds numbers produced by the weeds with the same rank. Finally, the maximum number of plant population of IWO was adjusted dynamically according to the number of nondominated solutions obtained during each iteration. Then the NSIWO approach was applied to the design of a Permanent Magnet Brushless Direct Current (PMBLDC) Motor of Underwater Unmanned Vehicle (UUV). The obtained results were compared with NSGA-II which is widely used in motor optimization. Numerical results in terms of convergence and spacing performance metrics indicates that the proposed multi-objective IWO scheme is capable of producing good solutions.

Keywords: brushless direct current motor, multi-objective optimization, fast nondominated sorting, invasive weed optimization, Pareto optimality.

1 Introduction

Optimization of a motor is a multi-objective optimization problem with several variables and constraints. In recent years the computational cost having been reduced dramatically, researchers all over the world are paying a considerable amount of attention towards bio-inspiration and bio-mimicry, for solving computationally expensive optimization problems. Many algorithms have been developed for optimization of electromagnetic devices, such as evolutionary algorithms[1-4] and swarm intelligence paradigms[5-8]. Invasive Weed Optimization (IWO) is a simple but powerful metaheuristic algorithm which is recently developed by A.R. Mehrabian and C. Lucas

[*] Corresponding author.

H. Sun et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 329, DOI: 10.1007/978-3-319-12286-1_9

in 2006[9]. IWO draws inspiration from the ecological process of weeds colonization and distribution and is capable of solving general multi-dimensional, linear and nonlinear optimization problems. At a fundamental level, the key difference between IWO and classical genetic algorithms is the way in which new samples are generated. In genetic algorithms, new samples are produced by some recombination of selected 'parent' solutions. In IWO, new samples (seeds) are generated by weeds and randomly dispersed in the neighborhood of the parent weed.

IWO has found successful applications in many practical single objective optimization problems [10-12]. In 2011, Debarati Kundu extended the basic IWO for tackling multi-objective optimization problems by using fuzzy dominance [13]. Then, Liu Xiao proposed a multi-objective IWO to optimize the model for weapon target assignment in 2013[14].

This paper proposed a new multi-objective optimization algorithm based on fast nondominated sorting [7] and Pareto-front for the design of PMBLDC motor. The concept of fast nondominated sorting was used to sort the promising candidate solutions, and each weed was allowed to produce seeds depended on its own rank. Moreover, crowding distance was used to solve the problem of bad distributed of Pareto-front. Finally, a dynamic maximum number of population was applied to competitive exclusion in IWO. With these improvements, the multi-objective evolution direction of invasive weed optimization algorithm was clearly pointed.

The proposed NSIWO approach was applied to a PMBLDC motor design of UUV. The comparison results indicated that the NSIWO could appear as a very promising candidate metaheuristic in the domain of multi-objective optimization, and is better than NSGA-II.

2 Multi-objective Nondominated Sorting Invasive Weed Optimization

The main framework of NSIWO is the same as that of IWO, while fast nondominated sorting approach, developed by Kalyanmoy Deb, is taken for finding the Pareto-front. In fast nondominated sorting approach, two entities are calculated for each solution: 1) domination count n_p , the number of solutions which dominate the solution p , and 2) S_p , a set of solutions that the solution p dominates. The computational complexity is $O(\mathrm{MN}^2)$ while the old nondominated sorting algorithm has a computational complexity of $O(\mathrm{MN}^3)$ [15].

IWO is a population-based meta-heuristic algorithm that mimics the colonizing behavior of weeds. The basic characteristic of a weed is that its population grows entirely or predominantly in a geographically specified area which can be substantially large or small. Initially, a finite number of seeds are dispread over the search area and every seed grows to a plant and produces seeds depending on its fitness, then the produced seeds are randomly dispread over the search area and grow to new plants. This process continues until maximum number of plants is reached. Finally, the plants with better fitness can survive and produce seeds, and others are eliminated. The process continues until the maximum iteration is reached and the plant with the best fitness is the closest to the optimal solution.

The main problem of multi-objective invasive weed optimization is the distribution of seeds produced by each weed. IWO approach proposes that the weeds with higher fitness would produce more seeds; therefore the weeds will constriction to the one with highest fitness. In multi-objective problem, the weeds which close to the Pareto front means higher “fitness”, and should naturally produce more seeds. Consequently, we proposed that the weeds were ranked by fast nondominated sorting approach, and the lower rank of weeds produced more seeds. This seeds distribution method will guarantee all the weeds growth to the Pareto front.

If the same rank of seeds produced the same number of seeds, the most nondominated solutions will crowd in a small region, and will get a bad distribution of Pareto-front. Therefore, we proposed that the number of seeds produced by a weed increased linearly from shortest crowding distance to the longest crowding distance. With this improvement, the sparse solutions of Pareto-front will produce more seeds to improve the distribution of solution

In IWO, the maximum number of plant population $P_{\max}$ is fixed. In the multi-objective IWO, the number of nondominated solutions was limited by the fixed $P_{\max}$, so we proposed a dynamic $P_{\max}$ that the value of $P_{\max}$ was adjusted according to the number of Pareto-front obtained by each iteration.

The implementation of NSIWO was based on following steps:

i) Initialize a finite number of weeds, using a generator of random solutions based on uniform distribution. Set the maximum iterations $iter_{\max}$, the initial maximum number of plant population $P_{\max\_\mathrm{ini}}$, the maximum and minimum number of seeds $S_{\max}$ and $S_{\min}$, the nonlinear modulation index n, the initial and final value of standard deviation $\sigma_{initial}$ and σ_{final};

ii) Sorting the population by using fast nondominated sorting approach and calculating the standardization crowding distance $\mathrm{d_i}$ of each weed. If the weeds of first rank N_{r1} exceeds $P_{\max}$, let $P_{\max} = N_{r1}$;

iii) Each weed was allowed to produce seeds depending on its own rank and crowding distance. The seeds of each weed was given by :

$$S_i = S_{\max} - (1 - \mathrm{d_i}) \frac{r_i(k)}{r_{\max}} (S_{\max} - S_{\min}) \tag{1}$$

Where $r_i(k)$ is the rank of weed, $r_{\max}$ is the maximum rank;

iv) The seeds of each weed were randomly distributed near to the parent weed with mean equal to zero but varying variance. The standard deviation of the random function was reduced from a previously defined initial value $\sigma_{initial}$ to a final value σ_{final} in every iteration according to equation 2.

$$\sigma_{iter} = \frac{(iter_{\max} - iter)^n}{(iter_{\max})^n}(\sigma_{initial} - \sigma_{final}) + \sigma_{final} \tag{2}$$

v) When the weed population exceeded the maximum number of plant population $P_{\max}$, the lower rank of weeds were allowed to survive, and upper rank of weeds are discarded;

vi) Continue until stopping criterion was met.

3 PMBLDC Motor Design Using NSIWO

3.1 Objective Function and Constraints

The optimization problem is the design of a PMBLDC motor for a UUV, which has 8 optimization variables and 4 constraints. The optimization parameter schematic diagram was shown in Fig. 1, and the optimization variable was shown in Table 1. Some other parameters determined by the material properties and manufacture technology such as the slot filling factor, the residual flux density of permanent magnet, the flux density on the knee point of B - H curve of stamping and so on, need to be confirmed beforehand.

The PMBLDC motor of UUVs requires higher efficiency and lower mass. Therefore, two objectives are defined as: f_1 represents the mass and f_2 represents the efficiency.

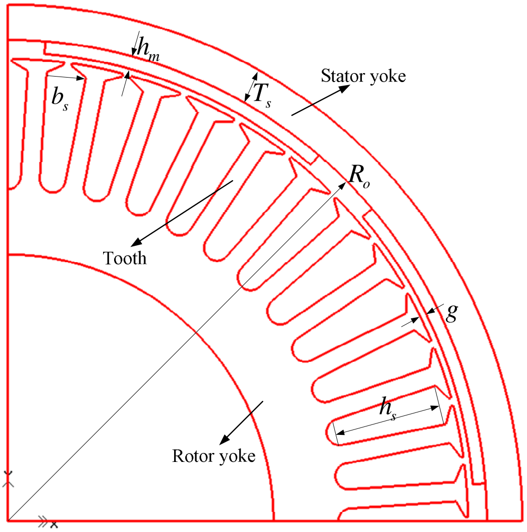


Fig. 1. Optimization parameters schematic diagram

Table 1. Optimization variables

Symbol	Variable	Lower bound	Upper bound
R_o (mm)	Inner radius of stator	80	250
h_m(mm)	Thickness of permanent magnet	2	10
g(mm)	Air gap	1	6
L_{st}(mm)	Thickness of lamination	100	200
$J(A/\mathrm{mm}^2$	Current density	4	6
$B_s(T)$	Magnetic flux density of stator yoke	1.3	1.6
$B_r(T)$	Magnetic flux density of rotor yoke	1.3	1.6
$B_t(T)$	Magnetic flux density of tooth	1.3	1.6

The constraints include: the supply voltage U_{dc} which is depended on the battery on UUV, the temperature T when the UUV is running, the minimum air gap $l_g^{\min}$, the ratio of slot depth and slot width S_{HB}. Therefore， the multi-objective problem could be written as:

$$\begin{cases}\min f_1 \\ \max f_2\end{cases} \tag{3}$$

Submitted to:

$$\begin{cases}U_{dc} = 200V \\ T < 120^\circ \\ 2 \le S_{HB} \le 5 \\ l_g^{\min} \ge 1mm\end{cases} \tag{4}$$

3.2 Mass and Efficiency of PMBLDC Motor

The magnetic field distribution in the airgap regions of a surface mounted permanent magnet BLDC motor can be presented by the following expression[16]:

$$B_{airgap}(\alpha, r) = B_{air-slotless}(\alpha, r)\tilde{\lambda}(\alpha, r) \tag{5}$$

Where $B_{air-slotless}(\alpha, r)$ is the magnetic flux density of airgap without considering the slot factor, and $\tilde{\lambda}(\alpha, r)$ is relative air gap permeance. Then, the magnetic flux of each pole is written as：

$$\phi_m = \int_0^{\frac{\pi}{p}} B_{airgap}(\alpha, r) d\alpha \tag{6}$$

Therefore the width of stator yoke, lamination tooth and rotor yoke can be respectively expressed as:

$$T_{stator-yoke} = \frac{0.5\phi_m}{B_s L_{st}} \tag{7}$$

$$T_{tooth} = B_{air-mean} \frac{2\pi R_i}{N_s B_t} \tag{8}$$

$$T_{rotor-yoke} = \frac{0.5\phi_m}{B_r k_j L_{st}} \tag{9}$$

Where k_j the stacking factor of the iron laminations is, $B_{air-mean}$ is the mean magnetic flux density of airgap. The turns per phase is deduced as：

$$W_\phi = \frac{7.5\alpha_p(U - 2\Delta U)}{pn_0\phi_m} \tag{10}$$

Where n_0 is the no-load speed, ΔU is the voltage drop of IGBT.

Then the mass is written as:

$$M = M_{rotor} + M_{stator} + M_{pm} + M_{copper} \tag{11}$$

The power loss including electrical, magnetic and mechanical power loss is expressed as[17]:

$$P_{loss} = 3I^2R_a + k_h B_{\max}^n f + 0.9078k_e B_{\max}^n f^2 + 0.96k_x B_{\max}^{1.5} f^{1.5} + 0.5\mu_f F_b d_i \omega_r \tag{12}$$

Where R_a is the resentence of armature; k_h , k_e and k_x are the coefficients of hysteresis, eddy current and excess eddy current losses; f is the frequency; $B_{\max}$ is the core maximum flux density ; n is the Steinmetz-constant; F_b is radial load of the bearing; d_i is the inner diameter of the bearing; μ_f is the friction coefficient of the bearing and ω_r rotational speed of the rotor.

4 Results

The NSGA-II was used to compare with NSIWO for the PMBLDC motor design. The control parameters for NSGA-II included: a population size of 100, crossover and mutation probabilities of 0.9 and 0.1, maximum iterations of100 . The control parameters for NSIWO were: initialize weeds $N_w = 100$ ， $iter_{\max} = 100$, $P_{\max\_ini} = 100$, $S_{\max} = 5$,

$S_{\min}=0$, $n=3$, $\sigma_{initial}=3$, $\sigma_{final}=0.001$.All the programs were run under Windows XP on a 2.3GHz Intel（R) Core(TM) 2 Duo CPU E6550 processor, with 2 GB of random access memory.

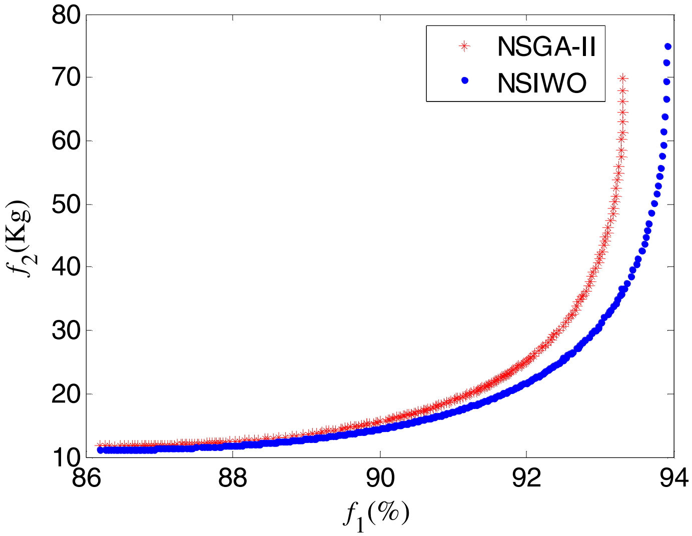


Fig. 2. Pareto-front points using NSIWO and NSGA- II

Table 1. Spacing and Euclidian Distances Indexes

Index	NSGA-II	NSIWO
Mean spacing (f_1, f_2)	2.2009	1.8551
Mean nomalized Euclidian distance (f_1, f_2) until the point of (100,0)	4.8267	4.7456
Pareto solutions	220	272
Mean computational time(minutes)	12	26

Fig. 2 shows the simulation results of NSIWO and NSGA-II. According to Fig.2, the nondominated solutions obtained by NSIWO dominate the solutions obtained by NSGA-II and with a good distribution and good spread compared with NSGA-II.

With the aim to investigate the mean spacing and mean Euclidian distances indexes for NSIWO and NSGA-II approaches, twenty independent simulation runs with different initial conditions were performed, and the results were shown in Table 2. According to Table 2, NSGA-II presented a smaller value of mean spacing between the objective functions f_1 and f_2, but NSIWO obtained better value in terms of mean normalized Euclidian distance between f_1 and f_2. In the aspect of optimization time, the NSIWO algorithm took a longer time compared with the NSGA-II algorithm. This is because although the NSGA-II and NSIWO have the same number of initial population and iteration, the total number of seeds produced by weeds in NSIWO was greater than the initial number of population, and the time complexity was increased.

In order to validate the optimization results, one of the solutions has been compared with the results of FEM. According to Fig.3, the presented flux density of airgap is in perfect agreement with Finite element method (FEM).

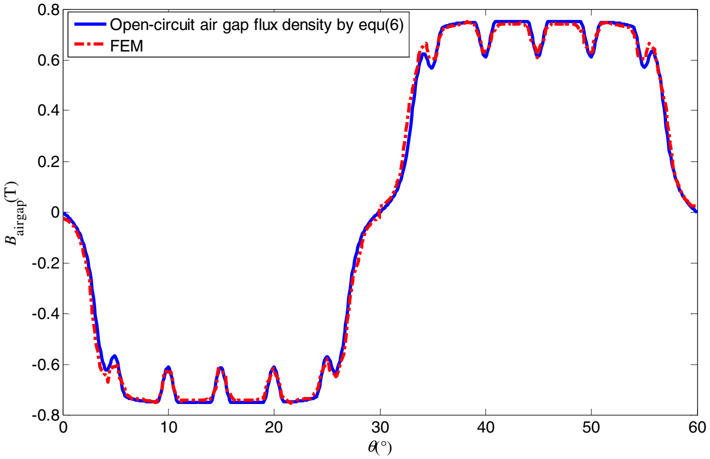


Fig. 3. Open-circuit magnetic flux distribution of the designed motor

5 Conclusions

In this paper, we presented a new multi-objective algorithm called nondominated sorting invasive weed optimization algorithm which was inspired in NSGA-Ⅱ and IWO. The fast nondominated sorting approach and crowding distance were used to solve the problem of which weeds should produce more seeds. In comparison with the NSGA-II algorithm, the solution sets of Pareto of both methods are competitive on a problem with high dimension and several constraints, but the nondominated solution obtained by NSIWO has a minor distance between the generated and Pareto-front. When the mean spacing and mean normalized Euclidian distance are used, the NSIWO outperforms NSGA-II. Numerical results indicate that the proposed multi-objective IWO scheme is capable of producing good solutions. Although the optimization algorithm is applied to the optimization design of PMBLDC motor, this method is also applicable to other types of motor design and multi-objective optimization problems.

References

1. Upadhyay, P.R., Rajagopal, K.R.: Genetic algorithm based design optimization of a permanent magnet brushless DC motor. Journal of Applied Physics 10, 10Q516–10Q516-3 (2005)

2. Yang, Y.P., Chiao, T.C.: Multi-objective optimal design of a high speed brushless DC motor. Electric Machines and Power Systems 28, 13–30 (2000)
3. Vaez-Zadeh, S., HassanpourIsfahani, A.: Multiobjective Design Optimization of Air-CoreLinear Permanent-Magnet Synchronous Motors forImproved Thrust and Low Magnet Consumption. IEEE Transactions on Magnetics 42, 446–452 (2006)
4. Chun, Y.D., Wakao, S., Kim, T.H., Jangand, K.B., Lee, J.: Multiobjective Design Optimization of Brushless Permanent Magnet Motor Using 3D Equivalent Magnetic Circuit Network Method. IEEE Transactions on Applied Superconductivity 14, 1910–1913 (2004)
5. dos Santos Coelho, L., Barbosa, L.Z., Lebensztajn, L.: Multi-objective Particle Swarm Approach for the Design of a Brushless DC Wheel Motor. IEEE Transactions on Magnetic 46, 2994–2997 (2010)
6. An, Y., Sun, C., Meng, Z., Che, D., Kong, Q., Cao, J.: Optimization Design of High Efficiency Permanent Magnet Spinning Motor with Hybrid Algorithm of PSO and Chaos. In: Proceeding of International Conference on Electrical Machines and Systems 2007, pp. 1778–1780 (2007)
7. Sakthivel, V.P., Bhuvaneswari, R., Subramanian, S.: Multi-objective parameter estimation of induction motor using particleswarm optimization. Engineering Applications of Artificial Intelligence 23, 302–312 (2010)
8. Duan, Y., Harley, R.G., Habetler, T.G.: Multi-objective Design Optimization of Surface Mount Permanent Magnet Machine with Particle Swarm Intelligence. In: IEEE Swarm Intelligence Symposium (2008)
9. Mehrabian, A.R., Lucas, C.: A novel numerical optimization algorithm inspired from weed colonization. Ecological Informatics 1, 355–366 (2006)
10. Mallahzadeh, A.R., Oraizi, H., Davoodi-Rad, Z.: Application of the invasive weed optimization technique for antenna configuration. Progress in Electromagnetics Research 79, 137–150 (2008)
11. Mallahzadeh, A.R., Es'haghi, S., Alipour, A.: Design of an E-Shaped Mimo Antenna Using IWO Algorithm for Wireless Application at 5.8 GHz. Progress in Electromagnetics Research 90, 187–203 (2009)
12. Mallahzadeh, A.R., Es'haghi, S., Hassani, H.R.: Compact U-array MIMO antenna designs using IWO algorithm. International Journal of RF and Microwave Computer-Aided Engineering 5, 568–576 (2009)
13. Kundu, D., Suresh, K., Ghosh, S.: Multi-objective optimization with artificial weed colonies. Information Sciences 181, 2441–2454 (2011)
14. Liu, X., Liu, Z., Hou, W., Xu, J.: Solving multiobjective optimization model for weapon target assignment by NRIWO algorithm. J. Huazhong Univ. of Sci.& Tech (Natural Science Edition) 41, 68–72 (2013)
15. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II. IEEE Transactions on Evolutions Computation 2, 182–197 (2002)
16. Zhu, Z.Q., Howe, D., Bolte, E., Ackermann, B.: Instantaneous magnetic field distribution in brushless permanent magnet DC motors, part I: open-circuit field. IEEE Transactions on Magnetics 1, 124–135 (1993)
17. Rahideh, A., Korakianitis, T., Ruiz, P., Keeble, T., Rothman, M.T.: Optimal brushless DC motor design using genetic algorithms. Journal of Magnetism and Magnetic Materials 322, 3680–3687 (2010)

An Evolutionary Approach to Tower of Hanoi Problem

Jie Li and Rui Shen

Department of Computer Engineering
Shandong Aerospace Electro-Technology Institute
Yantai, China
jie_yi_ehw@163.com

Abstract. The Tower of Hanoi problem is an ancient and interesting topic. In this paper, we presented an evolutionary algorithm approach for searching the solutions of the problem. We use a direct encoding and apply mutation only in the evolution. Experimental results are reported and show that the proposed method is capable of finding solutions for the problem of multiple pegs.

Keywords: Tower of Hanoi, Evolutionary approach, multiple pegs problem.

1 Introduction

The Tower of Hanoi problem, proposed by Lucas over a hundred years ago [1], is a well-known game that transferring a number of disks to a goal position. In the game, there exists three vertical pegs, and a player is given a certain number of disks (typically 3) of mutually different diameters place in small-on-large ordering on a peg. The task is to get from a given initial state to a target state by moving a single disk from the top of the peg to the top of another possible one, while obeying the following rules:

(1) Each time only one disk is moved;
(2) Only the topmost disk can be moved;
(3) At any moment, a disk cannot reside on a smaller one.

The Tower of Hanoi problem has been widely discussed in [2], [3], [4] and is being applied in many fields. In computer programming course, it is the most popular example for recursive programming [5], often comparing with the iterative solution, and showing that the recursive one is short and elegant. In psychology researches, the Tower of Hanoi is used as a tool to investigate how humans develop their problem solving ability [6], [7]. Many science and engineering activities require planning. Solving the Tower of Hanoi problem is a good example for modeling the process of planning, for instance, robot task plan [8], and unmanned vehicle route planning [9].

Work on this problem still goes on, studying properties of solution instances, as well as variants of the original problem. A natural extension of the original problem is obtained by adding pegs. In this work, we applied evolutionary algorithm to search adequate solutions for the Tower of Hanoi problem. Each particular moving is

H. Sun et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 329, DOI: 10.1007/978-3-319-12286-1_10

assigned by a specific integer number, which denotes a gene in EA. Thus, a series of disks moving steps can be described by a string of certain fixed length. All possible candidates are composed of such strings. The evolutionary algorithm checks the validation of the strings, and finally finds the solution. This work focuses on the problems of 3, 4 and 5 pegs. Experiment results show that our approach is able to solve the Tower of Hanoi problem of 3-peg with 5 disks, 4-peg with 8 disks and 5-peg with 9 disks.

The rest of the paper is organized as follows: Section 2 presents the detail of the non-determination approach. Section 3 describes the evolutionary algorithm applied in the work. Section 4 shows the experimental results. Section 5 gives the conclusion of the paper.

2 Proposed Method

2.1 Problem Description

The classic Tower of Hanoi problem has three pegs, denoted as A, B, C, and n (≥1) disks of different size. All disks initially rest on the source peg in a tower in small-on-large ordering, generally with the largest disk at the bottom, the second largest one above it, and so on, with the smallest one at the top. The objective is to transfer the tower from the source peg to the destination peg, with legal moves. The legal move is considered as the operation that can transfer the topmost disk from any peg to another without breaking the rules mentioned above.

Fig. 1 shows the initial and goal states of a 3-disk classical Tower of Hanoi problem. There are 3 disks, D1, D2, and D3 of increasing size. Initially, all the disks are on stake A, the source peg, while the target peg is stake C.

For the version of multi-peg Tower of Hanoi problems, the goal of the game is the same - moving the tower from the source to the destination.

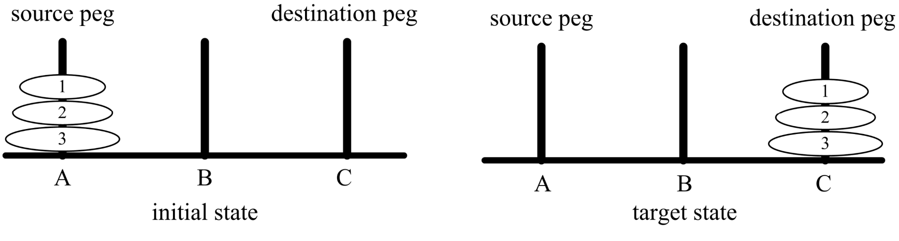


Fig. 1. The basic three-peg Tower of Hanoi Problem

2.2 Solution Encoding

Let M (n, p) denote the sequence of moves to solve the Tower of Hanoi problem with n (≥3) disks and p (≥3) pegs. And m_i denotes the ith move that transfer a disk from one peg to another. Then, a candidate plan for solving the Tower of Hanoi problem will be as follows:

$$M\,(n, p) \quad = \{\ m_0, m_1, m_2, \ldots ,m_{\mathrm{i}}, \ldots ,m_{\mathrm{k-1}}\ \}, \qquad (1)$$

where k is the length of the planning.

Each single operation in the plan is encoded as a gene, and the sequence of moves of a candidate planning M (n, p) is mapped as a combination of genes as the component of an individual, which the population is composed of.

In this paper, a direct way to represent an individual is used to encode each move by an integer, thus an individual is mapped as a string of integers. Take Fig.1 as an example. Let

1) 0 encodes the move from A to B
2) 1 encodes the move from B to A
3) 2 encodes the move from A to C
4) 3 encodes the move from C to A
5) 4 encodes the move from B to C
6) 5 encodes the move from C to B
7) 6 encodes No Action

The chromosome of a possible solution with a length of 7 for the problem can be as follows:

$$M\,(3, 3) = \{2, 0, 5, 2, 1, 4, 2\}.$$

This represents the sequence of operations that successfully transfer the three disks from A to C:

{A->C, A->B, C->B, A->C, B->A, B->C, A->C}

It is the simplest encoding way. However, the drawback is obvious: this encoding method do not guaranty all operations are valid in every possible state, and thus may result in an invalid solution that violates the three rules, even with the final goal state. For example, one possible sequence of moves could be as follows:

$$M\,(3, 3) = \{0, 0, 0, 4, 4, 4, 6\},$$

Which means a series of actions: {A->B, A->B, A->B, B->C, B->C, B->C, NA}.

The final result of these operations is correct. However, the repeat of moving disks from A to B introduces states with a larger disk reside on a smaller one, which breaks the 3rd rule. And thus, it is an illegal solution. So examination is necessary for solutions in the process of evaluation to ensure the validity of the final result.

Allowing the length of individual variable makes the evolutionary approach design more flexible but too complex. In this work, the length of individual is set with a specific fixed value for each evolution process.

2.3 Selection and Mutation

Selection. The initial population is generated randomly in the beginning of evolution. After evaluation, an elitism selection is employed to keep the best individual for generating the next population. If there finds an individual scores the same best fitness as the current best individual, the current best one will be replaced by the individual with a probability of 50%.

Mutation. Suppose the population contains S individuals. In a traditional way, a new population is generated by performing mutation operation on the best individual selected in the previous generation S-1 times in the beginning of each generation. In this paper, a hierarchical mutation mechanism is employed to create the new population. A new population, except the best individual, is divided into q groups. Each part contains the same number, say r, of individuals. In other words,

$$S-1=\sum_{i=1}^{q} r_i \tag{2}$$

The r individuals in the first group are directly generated from the best individual by mutation. Each single mutation on the best individual creates a new individual. Every gene of the parent has equal probability of being mutated. In each mutation operation, a new integer number that encodes a possible operation is randomly chosen to replace the previous one - the old gene.

For the second group, every individual in this group is produced from the one and only one individual in the first group. Since the two groups are of the same size, the one-to-one method can create the new group easily and quickly. Similar operation is taken to generate the 3rd group, and so on, until the generation of the whole new population has completed, as Fig.2 shows.

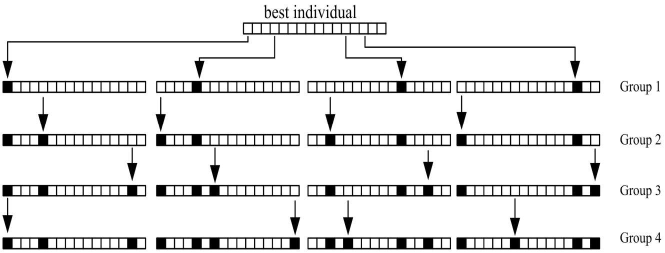


Fig. 2. Hierarchical mutation

In the example of Fig.2, the size of the population is S = 17. The rest individuals, excluding the best one, are separated into q = 4 groups, each group contains r = 4 individuals. An arrow line shows the offspring is produced by whom and a black block indicates a mutated gene. Note that the gene to be mutated is randomly selected without any additional rule, so the existence of duplication of individuals is allowed.

The intension of this mutation mechanism and the selection of q and r are trying to find a path towards the balance between exploitation and exploration, while hoping there will be a positive impact on the overall execution time of this approach.

Evaluation. The goal of the proposed approach is to find a solution that satisfies the following two conditions: first, the sequence of operations leads the system from the initial state to the goal state; second, no invalid operation is allowed in the solution.

The algorithm encourages individuals in the situation of less invalid moves in the sequence of operations, and later the first invalid move arises.

Therefore, the fitness function consists of three components: the score of the final state f_s, the number of invalid operations f_e, and the location of the first invalid move appears f_f.

The f_s evaluates the final state quality of a solution how well it fits to the goal state. The algorithm follows the operations from the beginning to the end of a solution to check its final state. The calculation of the score is related to the number of disks on the destination peg and the order of these disks. A better solution results in a higher fitness. The value of f_s depends on the number of pegs and disks.

While following the operations, the algorithm goes through every operation from the first move and checks if it is a valid operation. No matter what answer it is, the state is changed to what it will be, according to the behavior of the operation. If it is a valid operation, the values of f_e and ff remain unchanged. Otherwise, the value of f_e increases by 1. If it is the first invalid operation, the value of f_f is saved according to the location it appears, as follows:

$$f_f \;=\; max\_len - first\_error\_location \tag{3}$$

Where *max_len* is the length of the solution. And *first_error_location* is the location counted from the left side to the right side of the operation sequence. The later the first invalid operation appears, the less the f_f is. In the beginning of evaluation, the f_f is set to 0 initially. The fitness function is formulated as follows:

$$fitness = a \times f_s - b \times f_e - c \times f_f \tag{4}$$

Where *a*, *b* and *c* are weights. One can use them to emphasize one of the three factors by selecting adequate values, if it is necessary. In this paper, the three weights, *a*, *b* and *c* are all fixed as 1.

3 Evolutionary Algorithm

3.1 Algorithm Description

In this work, a modified CGP algorithm [10] is used as the evolutionary algorithm for searching the solution of the Tower of Hanoi problem. The basic procedure of the standard CGP algorithm is described as follows [11]:

1. Generate initial population
2. Evaluate fitness of genotypes in population
3. Promote fittest genotype to new population
4. Fill remaining places in the population with mutated versions of the parent
5. Return to step 2 until stopping criterion matched (reached the maximum generation number or found a valid solution)

Comparing with the conventional CGP, the modified version applies a hierarchical mutation mechanism, which enables the algorithm to search in a larger area with fewer amounts of genetic operations [12]. There is a slight difference in the evolutionary algorithm between this work and in [12]. The population size varies as the complexity of the problem increase. And no taboo strategy is used further, as can be seen from Fig.2.

3.2 Building Blocks

As tradition way goes, we choose a set of integer numbers as the building blocks for evolution. However, it is quite different from that of evolvable hardware. In evolvable hardware evolution, once the function set is selected, it remains unchanged no matter how complex the target circuit will be. Nevertheless, in the Tower of Hanoi problem the available type of operation grows as the member of pegs increases. The following equation describes the relation between the number of pegs and the number of available operations:

$$op = \mathrm{p}_m^2 = m \times (m-1) \tag{5}$$

Where m is the number of pegs, and op denotes the number of the available operations. Table 1 gives the building block set applied in this work.

For 3-peg Tower of Hanoi problem, the available building blocks in Table. I are those from No.1 to No.7. For 4-peg problem, the available range extends to No.13 and to No.21 for 5-peg problem. The table can be easily expanded as the peg number increases.

Table 1. Building block set

No.	Building block	Operation	No.	Building block	Operation
1	0	A->B	*12*	11	C->D
2	1	B->A	*13*	12	D->C
3	2	A->C	*14*	13	A->E
4	3	C->A	*15*	14	E->A
5	4	B->C	*16*	15	B->E
6	5	C->B	*17*	16	E->B
7	6	NA	*18*	17	C->E
8	7	A->D	*19*	18	E->C
9	8	D->A	*20*	19	D->E
10	9	B->D	*21*	20	E->D
11	10	D->B			

4 Experimental Results

To evaluate the proposed approach, a series of tests on the Tower of Hanoi problem with 3-peg, 4-peg and 5-peg was performed. Each experiment ran multiple times and the results are reported here.

4.1 Three-Peg Tower of Hanoi Problem

The minimum number of moves to reach the goal state has been proved to be 2^n-1 [13], where n is the number of disks. In some of the tests, we set the size of individuals to the length which is slightly longer than 2^n-1. In others, we set the size to the length of the optimal solution, 2^n-1. We performed 50 runs in each case of disks (from 3 disks to 5 disks). Table 2 shows the experimental results.

For 6-disk problem, the length of an optimal solution increases to 63. Ten runs for 6-disk problem were performed. However, no solution was found within 6 hours in the tests.

4.2 Four-Peg Tower of Hanoi Problem

For the number of pegs greater than 3, things become difference. As for the case of 4-peg Tower of Hanoi problem, the size of possible states for an n disks problem increases to 4^n. The Frame-Stewart algorithm offers a way to find presumed-optimal solution [14]. It has not been proved to be optimal yet, but most assume that the solutions of Frame and Stewart are correct.

According to the Frame-Stewart algorithm, the length of the presumed optimal solutions for 4-peg of 3, 4, 5, 6, 7, 8 disks are 5, 9, 13, 17, 25, 33 [15], respectively. We used these presumed results to define the individual length in parts of the tests.

4.3 Five-Peg Tower of Hanoi Problem

For 5-peg Tower of Hanoi problem, situation is similar. The length of the presumed optimal solutions for 5-peg of 5, 6, 7, 8, 9 disks are 11, 15, 19, 23, 27 [15], respectively, according to the Frame-Stewart algorithm. We used these presumed results to define the individual length in parts of the tests.

For the 5-peg of 9 disks problem, the presumed optimal solution length is 27. We set the length of individual to 35, and tried 5 runs to search the solutions. Only one valid solution with a length of 32 was found, and the time cost was 2205 seconds.

Table 2. Experimental results of 3-peg problem

Disks	Runs	Length setting of individual	Successful ratio	Average execution time (s)	Optimal solution length[1]
3	50	12	100%	0.137	7
	50	7	100%	0.85	7
4	50	24	100%	21.6	17
	50	15	100%	22.8	15
5	50	40	100%	1842.4	31
	50	31	100%	2419.2	31

"Optimal solution length" is the length of the remained moves in the best solution after removing the "NA" operations.

Table 3. Experimental results of 4-peg problem

Disks	Runs	Length setting of individual	Successful ratio	Average execution time (s)	Optimal solution length
3	30	8	100%	0	5
	30	5	100%	0.49	5
4	20	16	100%	29.0	10
	20	9	100%	37.2	9
5	20	28	100%	195.9	20
	20	13	80%	245.1	13
6	10	20	100%	39.3	17
	5	17	80%	386.5	17
7	10	30	100%	513.4	28
	5	25	100%	815.0	25
8	10	40	100%	2020.3	35
	5	33	80%	5632.2	33

Table 4. Experimental results of 5-peg problem

Disks	Runs	Length setting of individual	Successful ratio	Average execution time (s)	Optimal solution length
5	10	16	100%	2.1	14
	10	11	100%	36.8	11
6	10	20	100%	19.3	17
	5	15	100%	177.4	15
7	10	19	100%	264.2	19
8	5	23	100%	3829.0	23
9	5	35	20%	-	32

4.4 Summary and Discussion

The simple evolutionary method employed a fixed length strategy and applied only mutation as its evolutionary operator. The experimental results show that the proposed approach can evolve solutions for multiple pegs of the Tower of Hanoi problems. However, this method is not guaranteed to find a valid solution as the complexity of the problem increases. And the execution time it costs in this situation is longer than other deterministic algorithms.

One possible way to improve the performance of our approach is to adopt certain heuristic method to guide the behavior of the evolutionary algorithm. For example, to trace the state of the disks on the top of each pegs [16] and guide the selection of reasonable evolutionary operation. Of course, it will introduce additional memory and computation time cost.

In the process of evolution, some candidates go into a state where most of the disks reside on the target peg in correct ordering, while the largest disk and the second largest disk are left on other pegs. Before putting the largest and the second largest disks on the target peg, all the disks on the target peg should have to be moved away firstly. This situation may cause it more difficult to reach the goal state. Modification of the evaluation should be done to let the fitness function to assign more score to the behavior that arranges the largest and the second largest disks.

Also can be seen from the experimental results, reasonable selection for the length of solutions will be a help to reduce the evolving time. However, the relationship between the length and the execution time is unclear yet.

5 Conclusion

In this paper, a non-deterministic approach for evolving solutions of the Tower of Hanoi problem is presented. The proposed method uses a simple and direct encoding, and thus candidates with invalid operations are allowed during evolution. The algorithm applies mutation only in the evolution process. The results of experiments show that our approach is capable of finding valid solutions for some cases. However, as the complexity of the problem increases, the approach experiences difficulties in finding solutions within a certain time. Possible suggestions on improving the search performance are discussed. We also set the individual length to a number smaller than the one given by the Frame-Stewart algorithm and try to exam whether the EA can find a legal solution. But so far it is not succeed. Future works will focus on these issues.

References

1. Hinz, A.M.: The tower of Hanoi. Enseign. Math 35(2), 289–321 (1989)
2. Er, M.C.: A representation approach to the tower of Hanoi problem. The Computer Journal 25(4), 442–447 (1982)
3. Hinz, A.M.: Shortest paths between regular states of the tower of Hanoi. Information Sciences 63(1), 173–181 (1992)
4. Klavžar, S., Milutinović, U.: Graphs S (n, k) and a variant of the Tower of Hanoi problem. Czechoslovak Mathematical Journal 47(1), 95–104 (1997)
5. Hayes, P.J.: Discussion and correspondence A note on the Towers of Hanoi problem. The Computer Journal 20(3), 282–285 (1977)
6. Chi, M.T.H., Glaser, R.: Problem-solving ability. Learning Research and Development Center, University of Pittsburgh (1985)
7. Spitz, H.H., Webster, N.A., Borys, S.V.: Further studies of the Tower of Hanoi problem-solving performance of retarded young adults and nonretarded children. Developmental Psychology 18(6), 922 (1982)

8. Cambon, S., Gravot, F., Alami, R.: A robot task planner that merges symbolic and geometric reasoning. In: ECAI, vol. 16 (2004)
9. McDermott, P.L., Carolan, T.F., Gronowski, M.R.: Application of Worked Examples to Unmanned Vehicle Route Planning. In: The Interservice/Industry Training, Simulation & Education Conference (I/ITSEC), vol. 2012(1). National Training Systems Association (2012)
10. Li, J., Huang, S.: Evolving in extended hamming distance space: hierarchical mutation strategy and local learning principle for EHW. In: Kang, L., Liu, Y., Zeng, S. (eds.) ICES 2007. LNCS, vol. 4684, pp. 368–378. Springer, Heidelberg (2007)
11. Miller, J.F., Thomson, P.: Cartesian genetic programming. In: Poli, R., Banzhaf, W., Langdon, W.B., Miller, J., Nordin, P., Fogarty, T.C. (eds.) EuroGP 2000. LNCS, vol. 1802, pp. 121–132. Springer, Heidelberg (2000)
12. Li, J., Huang, S.: Adaptive salt-&-pepper noise removal: a function level evolution based approach. In: NASA/ESA Conference on Adaptive Hardware and Systems, AHS 2008. IEEE (2008)
13. Romik, D.: Shortest paths in the Tower of Hanoi graph and finite automata. SIAM Journal on Discrete Mathematics 20(3), 610–622 (2006)
14. Klavžar, S., Milutinović, U., Petr, C.: On the Frame–Stewart algorithm for the multi-peg Tower of Hanoi problem. Discrete Applied Mathematics 120(1), 141–157 (2002)
15. Houston, B., Masun, H.: Explorations in 4-peg Tower of Hanoi. Technical Report TR-04-10 (2004)
16. Klavzar, S., Milutinovic, U., Petr, C.: Combinatorics of topmost discs of multi-peg tower of Hanoi problem. Ars Combinatoria 59, 55–64 (2001)

Multipopulational Metaheuristic Approaches to Real-Parameter Optimization

Václav Snášel[1] and Pavel Krömer[1,2]

[1] Department of Computer Science,
VŠB Technical University of Ostrava,
Ostrava, Czech Republic
{vaclav.snasel,pavel.kromer}@vsb.cz
[2] Department of Computer and Electrical Engineering,
University of Alberta, Edmonton AB T6G 2V4, Canada
pavel.kromer@ualberta.ca

Abstract. Multipopulational metaheuristic methods have been used to solve a variety of problems. The use of multiple populations evolved in parallel and exchanging data according to a particular communication strategy is known to mitigate premature convergence, enlarge diversity of the populations, and generally improve the results obtained by the methods maintaining a sole panmictic population of candidate solutions. Moreover, multipopulational algorithms can be easily parallelized and efficiently accelerated by contemporary multicore and distributed architectures. In this work, we study two populational real-parameter optimization metaheuristics in a traditional and multipopulational configuration, and propose a new heterogeneous multipopulational approach. The usefulness of the new method is briefly evaluated on experiments with several well known test functions for real-parameter optimization.

Keywords: Particle Swarm Optimization, Differential Evolution, Multipopulational methods, Real-parameter Optimization.

1 Introduction

Multipopulational metaheuristic algorithms form a family of parallel metaheuristic (i.e. evolutionary, swarm-intelligent) methods that introduce into the metaheuristic search and optimization process additional level of parallelism by separating the population (swarm) of candidate solutions into multiple sub-populations (subswarms, islands) that are evolved separately [1,2,16]. The sub-populations in such a distributed parallel model are, in contrast to the master-slave (farming) model and cellular model of parallel populational metaheuristics, independent instances of the original metaheuristic algorithm. The islands are evolved independently and can exchange selected candidate solutions in a process called *migration*. However, the evolution on each island might be also completely isolated with no interaction among the sub-populations whatsoever.

In this work we study the performance of two popular populational metaheuristics for real-parameter optimization and compare the results obtained by

H. Sun et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 329, DOI: 10.1007/978-3-319-12286-1_11

their panmictic and multipopulational variants. Moreover, we propose a new heterogeneous distributed multipopulational metaheuristic approach that combines different populational metaheuristic algorithms. The performance of all investigated methods is compared on a set of well-known real-parameter optimization test functions.

The rest of this paper is organized in the following way: section 2 shortly describes the concept of parallel metaheuristics. Section 3 summarizes the metaheuristic optimization methods used in this work and section 3.4 introduces a new heterogeneous multipopulational approach. The method is evaluated in section 4 and conclusions are drawn in section 5.

2 Parallel Metaheuristics

The use of multiple sub-populations is a traditional extension to the populational metaheuristics that has been shown useful for both serial and parallel variants of the algorithms [1,2]. The distributed *island model* is a popular model for parallel multipopulational metaheuristics under which a number of sub-populations evolves side-by-side independently. The existence of multiple populations introduces next level of complexity and parallelism into the metaheuristic search and optimization process. The sub-populations on each island are usually initialized with different random values and tend to follow different search trajectory and explore different areas of the fitness landscape [16]. The algorithms executed on each island can be differently parametrized in order to exploit a variety of search strategies at once [1,16].

Communication strategy and data interchange topology is an important aspect of distributed multipopulational algorithms [2,7,16]. The *topology* defines logical links between the sub-populations [2], i.e the way candidate solutions migrate between the islands. Common topologies include *fully connected* and *ring* topology (see fig. 1). Other island model parameters include [2]: *migration rate*, i.e. the number of candidate solutions exchanged between islands; *migration period*, i.e. the number of generations between migrations; and the migrant *selection and replacement strategy*.

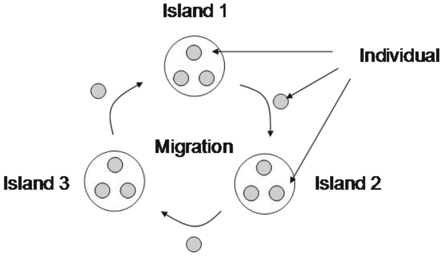


Fig. 1. Island model of populational metaheuristics with the *ring* topology [3]

The structure of the distributed multipopulational metaheuristic algorithms is suitable for parallelization on modern multicore and manycore platforms as well as in the environment of hybrid multi-CPU/GPU high performance computing (HPC) nodes and clusters of HPC nodes [2]. The structure is highly flexible and can be mapped to a variety of parallel architectures. Each island can be executed by a separate CPU thread or hosted on a separate compute node. The evolution of each island can be further parallelized and each sub-population can be executed using e.g. the master-slave model or the farming model.

In this paper, we compare the performance of the panmictic and multipopulational variant of two popular real-parameter optimization metaheuristics and propose a new heterogeneous multipopulational approach to real-parameter optimization.

3 Metaheuristic Algorithms for Real-Parameter Optimization

There is a number of metaheuristic populational algorithms that can be used for real parameter optimization. Differential evolution (DE) and particle swarm optimization (PSO) are among the most popular algorithms that use the real encoding of candidate solutions and simple operations to evolve them. In this section, we briefly summarize the basic principles of both algorithms and outline the proposed heterogeneous multipopulational approach.

3.1 Differential Evolution

The DE is a versatile and easy to use stochastic evolutionary optimization algorithm [12]. It is a population-based optimizer that evolves a population of real encoded vectors representing the solutions to given problem. The DE was introduced by Storn and Price in 1995 [13,14] and it quickly became a popular alternative to the more traditional types of evolutionary algorithms. It evolves a population of candidate solutions by iterative modification of candidate solutions by the application of the differential mutation and crossover [12]. In each iteration, so called trial vectors are created from current population by the differential mutation and further modified by various types of crossover operator. At the end, the trial vectors compete with existing candidate solutions for survival in the population.

The DE starts with an initial population of N real-valued vectors. The vectors are initialized with real values either randomly or so, that they are evenly spread over the problem space. The latter initialization leads to better results of the optimization [12].

During the optimization, the DE generates new vectors that are scaled perturbations of existing population vectors. The algorithm perturbs selected base vectors with the scaled difference of two (or more) other population vectors in order to produce the trial vectors. The trial vectors compete with members of the current population with the same index called the target vectors. If a trial vector represents a better solution than the corresponding target vector, it takes its place in the population [12].

There are two most significant parameters of the DE [12]. The scaling factor $F \in [0, \infty]$ controls the rate at which the population evolves and the crossover probability $C \in [0, 1]$ determines the ratio of bits that are transferred to the trial vector from its opponent. The size of the population and the choice of operators are another important parameters of the optimization process.

The basic operations of the classic DE can be summarized using the following formulas [12]: the random initialization of the ith vector with N parameters is defined by

$$x_i[j] = rand(b_j^L, b_j^U), \quad j \in \{0, \ldots, N-1\} \tag{1}$$

where b_j^L is the lower bound of jth parameter, b_j^U is the upper bound of jth parameter and $rand(a, b)$ is a function generating a random number from the range $[a, b]$. A simple form of the differential mutation is given by

$$v_i^t = v_{r1} + F(v_{r2} - v_{r3}) \tag{2}$$

where F is the scaling factor and v_{r1}, v_{r2} and v_{r3} are three random vectors from the population. The vector v_{r1} is the base vector, v_{r2} and v_{r3} are the difference vectors, and the ith vector in the population is the target vector. It is required that $i \neq r1 \neq r2 \neq r3$.

The uniform crossover that combines the target vector with the trial vector is given by

$$l = rand(0, N-1) \tag{3}$$

$$v_i^t[m] = \begin{cases} v_i^t[m] & \text{if } (rand(0,1) < C) \text{ or } m = l \\ x_i[m] & \end{cases} \tag{4}$$

for each $m \in \{1, \ldots, N\}$. The uniform crossover replaces with probability $1 - C$ the parameters in v_i^t by the parameters from the target vector x_i. The outline of the classic DE according to [6,12] is summarized in Algorithm 1.

The DE is a successful evolutionary algorithm designed for continuous parameter optimization driven by the idea of scaled vector differentials. That makes it an interesting alternative to the wide spread genetic algorithms that are designed to work primarily with discrete encoding of the candidate solutions. As well as GA, it represents a highly parallel population based stochastic search meta-heuristic. In contrast to the GA, the differential evolution uses the real encoding of candidate solutions and different operations to evolve the population. It results in different search strategy and different directions found by DE when crawling a fitness landscape of the problem domain.

3.2 Particle Swarm Optimization

The PSO algorithm is a global population-based search and optimization algorithm based on the simulation of swarming behavior of birds within a flock,

```
Initialize the population $P$ consisting of $M$ vectors using eq. (1);
Evaluate an objective function ranking the vectors in the population;
while Termination criteria not satisfied do
    for $i \in \{1, \dots, M\}$ do
        Differential mutation: Create trial vector $v_i^t$ according to eq. (2);
        Validate the range of coordinates of $v_i^t$. Optionally adjust coordinates of
        $v_i^t$ so, that $v_i^t$ is valid solution to given problem;
        Perform uniform crossover. Select randomly one parameter $l$ in $v_i^t$ and
        modify the trial vector using eq. (3);
        Evaluate the trial vector.;
        if trial vector $v_i^t$ represent a better solution than population vector $v^i$
        then
            add $v_i^t$ to $P^{t+1}$
        else
            add $v^i$ to $P^{t+1}$
        end
    end
end
```

Algorithm 1: A summary of classic Differential Evolution

schools of fish and even human social behavior [4,6,8]. PSO uses a population of motile candidate particles characterized by their position x_i and velocity v_i inside the $n-$dimensional search space they collectively explore. Each particle remembers the best position (in terms of fitness function) it visited y_i and knows the best position discovered so far by the whole swarm $\bar{y}$. In each iteration, the velocity of particle i is updated according to [6]:

$$v_i^{t+1} = v_i^t + c_1 r_1^t (y_i - x_i^t) + c_2 r_2^r (\bar{y}^t - x_i^t) \tag{5}$$

where c_1 and c_2 are positive acceleration constants and r_1 and r_2 are vectors of random values sampled from uniform distribution. Vector y_i^t represents the best position known to particle i in iteration t and vector $\bar{y}^t$ is the best position visited by the swarm at time t.

The position of particle i is updated by [6]:

$$x_i^{t+1} = x_i^t + v_i^{t+1} \tag{6}$$

The basic (*gbest*) PSO according to [6,8] is summarized in Algorithm 2.

PSO is useful for dealing with problems in which the solution can be represented as a point or surface in an $n-$dimensional space. Candidate solutions (particles) are placed in this space and provided with an initial (random) velocity. Particles then move through the solution space and are evaluated using some fitness function after each iteration. Over time, particles are accelerated towards those locations in the problem space which have better fitness values.

The *gbest* PSO yields good results in some application domains. Besides that, there is a number of alternative PSO versions including self-tuning PSO, niching

```
Create population of $M$ particles with random position and velocity;
Evaluate an objective function $f$ ranking the particles in the population;
while Termination criteria not satisfied do
    for $i \in \{1, \ldots, M\}$ do
        Set personal and global best position:
        if $f(x_i) < f(y_i)$ then
            $y_i = x_i$
        end
        if $f(x_i) < f(\bar{y})$ then
            $\bar{y} = x_i$
        end
        Update velocity of $i$ by (5) and position of $i$ by (6);
    end
end
```

Algorithm 2: Summary of *gbest* PSO

PSO, and multiple-swarm PSO that were developed in order to improve the algorithm or to solve difficult problems [4,6].

3.3 Multipopulational DE and PSO

Multipopulational DE [9,11] and multipopulational (multiswarm) PSO [10,17,5] represent straightforward as well as more sophisticated applications of the principles of parallel metaheuristics to the DE and PSO algorithms respectively. In this work, we use for the sake of comparison the island model applied to the traditional differential evolution (mDE) and particle swarm optimization (mPSO) without further extensions such as quantum particles [5], self-adaptation [17] etc. We compare them with the panmictic variants of the algorithms and with a new heterogeneous multipopulational approach.

3.4 Heterogeneous Multipopulational Approach to Real-Parameter Optimization

One of the main motivations behind the multipopulational approaches is the improvement of the diversity and the exploration of different trajectories on the fitness landscape. Different initialization and parametrization of the algorithms on different islands can contribute to different high-level search strategies (the preference of exploitation over exploration etc.) performed by each sub-population. A multipopulational approach placing not only different sub-populations but also *different populational metaheuristics* on different islands might strengthen these effects and eventually contribute to better optimization results.

The heterogeneous multipopulational approach (mHA) to real-parameter optimization is defined in the following way: no each island $i \in \{i_0, \ldots, i_n\}$ assign one of the *compatible* metaheuristic optimization algorithms $a_i \in \{a_0, \ldots, a_k\}$. Then perform the traditional multipopulational evolution. Two metaheuristic algorithms are said to be *compatible* if the representation of the candidate solutions allows seamless migration between their populations.

The combination of fundamentally different search strategies represented by the different metaheuristic algorithms and the exchange of the solutions found by these algorithms should contribute to better global optimization results. It should be also noted that the combination of several distinct algorithms in a multipopulational fashion can be used for any problem domain.

4 Experiments

A series of computational experiments was performed in order to compare the results of the panmictic and multipopulational PSO and DE with the heterogeneous multipopulational approach.

Several well known and widely used real-parameter optimization test functions [15] were selected as test problems. The functions were used as test problems with high computational requirements and known optima. The test functions used in this study are summarized in table 1. DE, PSO, mDE, mPSO, and mHA were implemented from scratch in C++ and used to search for the optima of the test functions with the dimension 50. The search was terminated after 500 generations of each algorithm.

Table 1. Test functions

Function name	Definition	Parameter range
Ackley	(7)	[-30, 30]
De Jong	(8)	[-5.12, 5.12]
Griewank	(9)	[-600, 600]
Rastrigin	(10)	[-5.12, 5.12]
Rosenbrock	(11)	[-30, 30]
Schwefel	(12)	[-500, 500]

$$f_1(\mathbf{x}) = -20 \cdot \exp\left(-0.2\sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} x_i^2}\right) - exp\left(\frac{1}{n}\sum_{i=1}^{n}\cos(2\pi x_i)\right) + 20 + e \tag{7}$$

where n is problem dimension, $\mathbf{x} = (x_1, \ldots, x_n)$ is parameter vector, $e \approx 2.71828$ is Euler's number, and $\exp(a) = e^a$ is exponential function.

$$f_2(\mathbf{x}) = \sum_{i=1}^{n} x_i^2 \tag{8}$$

$$f_3(\mathbf{x}) = 1 + \frac{1}{4000}\sum_{i=1}^{n} x_i^2 - \prod_{i=1}^{n} \cos\left(\frac{x_i}{\sqrt{i}}\right) \tag{9}$$

$$f_4(\mathbf{x}) = 10 \cdot n + \sum_{i=1}^{n} \left(x_i^2 - 10 \cdot \cos(2\pi x_i)\right) \tag{10}$$

$$f_5(\mathbf{x}) = \sum_{i=1}^{n-1} \left(100 \cdot (x_{i+1} - x_i^2)^2 + (1 - x_i)^2\right) \tag{11}$$

$$f_6(\mathbf{x}) = \sum_{i=1}^{n} -x_i^2 \cdot \sin(\sqrt{|x_i|}) \tag{12}$$

The parameters of all algorithms were selected on the basis of previous experience and best practices. The DE algorithm was */DE/rand/1* with $F = 0.9$ and $C = 0.9$ and the *gbest* PSO used local and global weights $c_1 = c_2 = 1.49445$. The number of sub-populations (islands) was set to 4, migration rate to 2, and migration period to 20 generations. The multipopulational variants used the ring topology and the size of the populations was fixed to 400 for DE and PSO and to 100 for each island of the mDE, mPSO, and mHA respectively so that the execution of each algorithm involved the same number of fitness function evaluations. Because the metaheuristics are stochastic, all experiments were repeated 30 times for each configuration and the results, presented in table 2, are averages of the 30 independent runs. The best results are *emphasized* (the lower the better).

Table 2. Average results

	Algorithm				
Function	DE	PSO	mDE	mPSO	mHA
Ackley	0.094705	0.664699	0.185931	0.00269948	*0.000709818*
Rosenbrock	276.571	104.279	400.807	107.684	*102.816*
Schwefel	9047.76	4494.65	5225.81	4465.01	*3950.77*
Rastrigin	149.821	135.711	*42.3665*	122.936	49.7072
De Jong	2.1089e-05	5.72866e-08	0.000103227	1.13755e-07	*4.18413e-08*
Griewank	0.00018074	1.96678e-07	0.000354401	3.90547e-07	*1.4365e-07*

The table shows that the heterogeneous mHA algorithm delivered best average result for 5 out of 6 test functions. The mDE algorithm outperformed mHA for Rastrigin's function. The results do not indicate that the use of mDE would bring any significant improvement over the use of DE. DE was, for some test functions, better than mDE and vice versa for some other test functions. The same was observed for the relationship between PSO and mPSO.

5 Conclusions

This study compared the performance of panmictic and multipopulational versions of two metaheuristic algorithms for real-parameter optimization. Moreover, it proposed a new heterogeneous multipopulational approach combining multiple distinct algorithms and multiple high-level search and optimization strategies into a single optimization approach.

The algorithms were tested on a set of well-known real-parameter optimization functions and the results show that the heterogeneous approach is a valid alternative to more traditional multipopulational algorithms as it obtained best average results for 5 out of 6 test functions. However, the multipopulational algorithms did not deliver significantly better results than their panmictic alternatives for the test functions employed in this study.

Acknowledgments. This work was partially supported by the Grant of SGS No. SP2014/110, VŠB - Technical University of Ostrava, Czech Republic, by the European Regional Development Fund in the IT4Innova-tions Centre of Excellence project (CZ.1.05/1.1. 00/02.0070) and by the Bio-Inspired Methods: research, development and knowledge transfer project, reg. no. CZ.1.07/2.3.00/20.0073 funded by Operational Programme Education for Competitiveness, co-financed by ESF and state budget of the Czech Republic.

References

1. Alba, E., Talbi, E.G., Luque, G., Melab, N.: Metaheuristics and Parallelism, pp. 79–103. John Wiley & Sons, Inc. (2005)
2. Alba, E., Luque, G., Nesmachnow, S.: Parallel metaheuristics: recent advances and new trends. International Transactions in Operational Research 20(1), 1–48 (2013)
3. Ando, J., Nagao, T.: Fast evolutionary image processing using multi-gpus (December 01, 2009)
4. Clerc, M.: Particle Swarm Optimization. ISTE, Wiley (2010)
5. Duan, Q., Wu, R., Dong, J.: Multiple swarms immune clonal quantum-behaved particle swarm optimization algorithm and the wavelet in the application of forecasting foundation settlement. In: 2010 2nd International Asia Conference on Informatics in Control, Automation and Robotics (CAR), vol. 3, pp. 109–112 (March 2010)
6. Engelbrecht, A.: Computational Intelligence: An Introduction, 2nd edn. Wiley, New York (2007)
7. Guan, W., Szeto, K.Y.: Topological effects on the performance of island model of parallel genetic algorithm. In: Rojas, I., Joya, G., Cabestany, J. (eds.) IWANN 2013, Part II. LNCS, vol. 7903, pp. 11–19. Springer, Heidelberg (2013)
8. Kennedy, J., Eberhart, R.: Particle swarm optimization. In: Proceedings of the IEEE International Conference on Neural Networks 1995, vol. 4, pp. 1942–1948 (1995)
9. Kundu, R., Mukherjee, R., Debchoudhury, S.: Multipopulation based differential evolution with self exploitation strategy. In: Panigrahi, B.K., Das, S., Suganthan, P.N., Nanda, P.K. (eds.) SEMCCO 2012. LNCS, vol. 7677, pp. 267–275. Springer, Heidelberg (2012)

10. Liang, J.J., Suganthan, P.: Dynamic multi-swarm particle swarm optimizer with local search. In: The 2005 IEEE Congress on Evolutionary Computation 2005, vol. 1, pp. 522–528 (September 2005)
11. Novoa-Hernández, P., Corona, C.C., Pelta, D.A.: Self-adaptive, multipopulation differential evolution in dynamic environments. Soft Computing 17(10), 1861–1881 (2013)
12. Price, K.V., Storn, R.M., Lampinen, J.A.: Differential Evolution A Practical Approach to Global Optimization. Natural Computing Series. Springer, Berlin (2005)
13. Storn, R.: Differential evolution design of an IIR-filter. In: Proceeding of the IEEE Conference on Evolutionary Computation, ICEC, pp. 268–273. IEEE Press (1996)
14. Storn, R., Price, K.: Differential Evolution- A Simple and Efficient Adaptive Scheme for Global Optimization over Continuous Spaces. Tech. rep. (1995)
15. Suganthan, P.N., Hansen, N., Liang, J.J., Deb, K., Chen, Y.P., Auger, A., Tiwari, S.: Problem definitions and evaluation criteria for the CEC 2005 Special Session on Real Parameter Optimization. Tech. rep., Nanyang Technological University (2005)
16. Whitley, D., Rana, S., Heckendorn, R.B.: The island model genetic algorithm: On separability, population size and convergence. Journal of Computing and Information Technology 7, 33–47 (1998)
17. Zhang, J., Ding, X.: A multi-swarm self-adaptive and cooperative particle swarm optimization. Engineering Applications of Artificial Intelligence 24(6), 958–967 (2011)

Evolved Bat Algorithm for Solving the Economic Load Dispatch Problem

Thi-Kien Dao[1], Tien-Szu Pan[1], Trong-The Nguyen[2], and Shu-Chuan Chu[3]

[1] Department of Electronics Engineering,
National Kaohsiung University of Applied Sciences, Taiwan
[2] Faculty of Information Technology, Haiphong Private University, Vietnam
[3] School of Computer Science, Engineering and Mathematics,
Flinders University, Australia
vnthe@hpu.edu.vn

Abstract. Economic Load Dispatch (ELD) is one of the important optimization tasks, which provides an economic condition for the power systems. In this paper, Evolved Bat Algorithm (EBA) as an evolutionary based approach is presented to solve the constraint economic load dispatched problem of thermal plants. The output generating power for all the power-generation units can be determined by the optimal technique for transmission losses, power balance and generation capacity, so that the total constraint cost function is minimized. A piecewise quadratic function is used to show the fuel cost equation of each generation unit, and the B-coefficient matrix is used to represent transmission losses. The systems with six units and fifteen units of thermal plants are used to test the demonstration of the solution quality and computation efficiency of the feasibility of the application of the Evolved Bat Algorithm for ELD. The experimental results compared with the genetic algorithm (GA) method for ELD, and with the particle swarm optimization (PSO) method for ELD, show that the applied EBA method for ELD can provide the higher efficiency and accuracy.

Keywords: Evolved Bat Algorithm, Economic Load Dispatch, Particle Swarm Optimization, Genetic Algorithm.

1 Introduction

Economic load dispatch problem is defined as the method of determining the optimal combination of power outputs for all generating units. The total fuel cost of thermal power plants needs to minimize while satisfying load demand and operating constraints of a power system. This leads the economic load dispatch to be a large-scale non-linear constrained optimization problem. The numerical programming based techniques such as lambda iteration and gradient-based methods could be used to solve this problem [1-3]. In these methods, a single quadratic polynomial was used to represent approximately the cost function of each generator. In that case, the fuel cost curves should be piece-wise linear and monotonically increasing to determine the global optimal solution. These techniques offered good results but when the search

H. Sun et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 329, DOI: 10.1007/978-3-319-12286-1_12

space was non-linear and it had discontinuities. They became very complicated with a slow convergence ratio and were not always seeking to the feasible solution. The ELD problem of finding the global optimal solution became the challenging.

Another numerical method was needed to cope with these difficulties, especially those with high-speed search to the optimal and not being trapped in local minima. Dynamic programming (DP) method [4] is one of the approaches to solve the non-linear and discontinuous ED problem, but it suffers from the problem of curse of dimensionality or local optimality.

Recently, many swarm intelligence algorithms have been developed and been prosperously used to solve optimization of this problem. For instance, particle swarm optimization (PSO) techniques have victoriously been used to solve the ELD problems and various power system problems [5] [6]; genetic algorithm (GA) [7] [8], and simulated annealing (SA) [9] [10] have proved to be very effective in solving nonlinear ELD problems without any restriction on the shape of the cost curves. Although these heuristic methods do not always guarantee discovering the globally optimal solution in finite time, they often provide a fast and reasonable solution. They can be a promising answer to overcome the above-mentioned drawbacks.

In this paper, a new swarm intelligence algorithm known as Evolved Bat algorithm is used to solve the economic load dispatch problem. Two cases with six units and fifteen units thermal power system are tested and compared with other approaches and found that the applied EBA method for ELD can provide the higher effect and accuracy and be promising method.

The rest of this paper is organized as follows: a brief description of the ELD problem associated with its mathematical formulation in session 2; a short review of EBA is given in session 3; details the proposed procedures with cases study is presented in session 4; the experimental results and the comparison of the applied EBA method for ELD with PSO and GA methods for ELD are discussed in session 5; finally, the conclusion is drawn in session 6.

2 Economic Load Dispatch Problem

The economic load dispatch problems (ELD) can be formulated as a nonlinear constrained problem [1, 3, 11]. The convex ELD problem assumes quadratic cost function along with system power demand and operational limit constraints. The objective of ELD problems is to minimize the fuel cost of committed generators (units) subjected to operating constraints. Practically, the economic power dispatch problem is usually formulated as: Minimize equation (1):

$$F_T = \sum_{i=1}^{n} F_i(P_i) \tag{1}$$

Subject to

$$\sum_{i=1}^{n} P_i = P_D + P_L \tag{2}$$

$$P_i^{Min} \leq P_i \leq P_i^{Max} \quad i = 1,2..n \tag{3}$$

where, F_T is the total fuel cost, n is the number of units, *Fi* and *Pi* are the cost function and the real power output of *i*th unit respectively, P_D is the total demand, P_L is the transmission loss, P_i^{Min} and P_i^{Max} and are the lower and upper bounds of the *i*th unit respectively. The equality constraint, Eqn.(2) states that the total generated power should be balanced by transmission losses and power consumption while Eqn.(3) denoting unit's operation constraints. Traditionally, the fuel cost of a generator is usually defined by a single quadratic cost function,

$$F_i(P_i) = \gamma_i P_i^2 + \beta_i P_i + \alpha_i \tag{4}$$

where, *αi, βi,* and *γi* are cost coefficients of the *i-th* unit. One common practice for including the effect of transmission losses is to express the total transmission loss as a quadratic function of the generator power outputs in one of the following forms[1]:

$$P_L = \sum_{i=1}^{n} \sum_{j=1}^{n} p_i B_{ij} p_j \tag{5}$$

Kron's loss formula:

$$P_L = \sum_{i=1}^{n} \sum_{j=1}^{n} p_i B_{ij} p_j + \sum_{j=1}^{n} B_{0j} p_j + B_{00} \tag{6}$$

where B_{ij} is called the loss coefficients.
B-matrix loss formula:

$$P_L = P^T B_P + P^T B_0 + B_{00} \tag{7}$$

where, *P* denotes the real power output of the committed units in vector form, and *B*, B_0 and B_{00} are loss coefficients in matrix, vector and scalar respectively, which are assumed to be constant, and reasonable accuracy can be achieved when the actual operating conditions are close to the base case where the *B*-coefficients were derived.

In the summary, the objective of economic power dispatch optimization is to minimize F_T subject to the constraints equation (2) and (3).

3 Meta-heuristic Evolved Bat-Inspired Algorithm

Evolved Bat Algorithm (EBA) is a new method in the branch of swarm intelligence for solving numerical optimization problems [12]. The EBA was proposed based on the framework of the original bat algorithm [13]. It was developed swarm intelligence algorithm inspired by bat echolocation in the natural world. The computational speed of the EBA is fast because its structure is designed with simple and light computations. In EBA, the general characteristics of whole species of bat were considered and the behavior of bats was reanalyzed to redefine the corresponding operation of the bats' behaviors. The sound speed spreads in the air was focused on as a major variable of determining for movement of the artificial agent because of all bats utilize the echolocation to detect their prey. The different between EBA and the original BA is the movement of the bat and the random walk process. The step size of the movement of the artificial agent in the solution space is determined by the chosen medium for spreading sound waves. Because of the direct influence on the search result is the step size. So the air of natural environment where bats live is chosen for the medium, that the sound speed spreads is *340* meter per second. In the active sonar system, the distance between the sound wave source and the target, which bounds the wave back, was defined by equation (8):

$$D = \frac{v \times \Delta T}{2} \tag{8}$$

where *D* denotes the distance; *v* is the sound speed; ΔT means the time difference between sending the sound wave and receiving the echo. The sound speed in the air is used to be the value of *v*. The unit of *v* is meter per second to kilometer per second. Thus, equation (8) can be reformed into equation (9):

$$D = \frac{340}{2}(m/sec.) \cdot \Delta T(sec.) = \frac{0.34}{2}(km/sec.) \cdot \Delta T(sec.) = 0.17\Delta T \tag{9}$$

In our experiments, the random number in the range of [-1, 1] is used to denote ΔT. The negative part of ΔT comes from the moving direction in the coordinate. ΔT is given with a negative value when the transmission direction of the sound wave is opposite to the axis of the coordinate. The movement of the bat in EBA is defined by equation (10):

$$x_i^t = x_i^{t-1} + D \tag{10}$$

where x_i^t indicates the coordinate of the *i-th* artificial agent; and t is the iteration number. In addition, if a bat moves into the random walk process, its location will be updated by equation (11):

$$x_i^{tR} = \beta \cdot (x_{best} - x_i^t) \tag{11}$$

where β is a random number in the range $\beta \in [0,1]$; x_{best} indicates the coordinate of the near best solution, that found so far overall artificial agents; x_i^{tR} represents the new coordinates of the artificial agent after the operation of the random walk process.

4 Proposed EBA for Solving the Economic Dispatch Problem

In this secession, the economic dispatch problem with the generator constraints as linear equality and inequality constraints and the transmission loss would be solved by applying the new scheme of Evolved Bat algorithm. The constraints relating to the ELD problem has to be handled with Evolved Bat algorithm. The search space in constrained optimization problems consists of two kinds of points: feasible and unfeasible. The feasible points satisfy all the constraints, while the unfeasible points violate at least one of them. Therefore, the solution or set of solutions obtained as the final result of an optimization method must necessarily be feasible, i.e., they must satisfy all constraints. The methods based on the use of penalty functions are usually employed to treat constrained optimization problems [14, 15].

A constrained problem can be transformed into an unconstrained one by penalizing the constraints and building a single objective function, which in turn is minimized using an unconstrained optimization algorithm. Several methods have been proposed to handle constraints in optimization problems [16], such as penalty- based methods, methods that preserve the feasibility of solutions, methods that clearly distinguish between feasible and unfeasible solutions and hybrid methods. When optimization algorithms are used for constrained optimization problems, it is common to handle constraints using concepts of penalty functions which penalize unfeasible solutions, i.e., one attempt to solve an unconstrained problem in the search space *solution* using a modified fitness function *f*. In this work, penalty function is applied for the fitness function of minimizing as following equation (12):

$$Min\ f = \begin{cases} f(P_i), & \text{if } P_i \in F \\ f(P_i) + penalty(P_i), & \text{otherwise} \end{cases} \tag{12}$$

where, penalty *(Pi)* is zero and no constraint is violated; otherwise it is positive. The penalty function is usually based on a distance measured to the nearest solution in the feasible region *F* or to the effort to repair the solution.

$$Min\ f = \sum_{i=1}^{n} F_i(P_i) + q_1\left(\sum_{i=1}^{n} P_i - P_L - P_D\right)^2 + q_2 \sum_{j=1}^{n} V_j \tag{13}$$

where q1 and q2 are penalty factors, positive constants associate with the power balance and prohibited zones constraints, respectively. These penalty factors were tuned empirically and their values are $q_1 = 1000$ and $q_2 = 1$ in the studied cases in this work. The V_j is expressed as follows:

$$V_j = \begin{cases} 1, & \text{if } P_j \text{ violates the prohibited zones} \\ 0, & \text{otherwise} \end{cases} \tag{14}$$

The process of EBA for ELD is depicted as follows:

*Step*1. Initialization: the artificial agents are spread throughout the solution space by randomly assigning the coordinates to them.
*Step*2. Movement: the artificial agents are moved by equations (9), (10). A random number is generated and then it is checked whether it is greater than the fixed pulse emission rate. If the result is positive, the artificial agent is moved using the random walk process, as defined by (11).
*Step*3. Evaluation: the fitness of the artificial agents is calculated by the fitness function in equation (13) and updated to the stored near best solution.
*Step*4. Termination: the termination conditions are checked to decide whether to go back to step 2 or terminate the program and output the near best solution.

5 Experimental Results

The proposed scheme is implemented for six units and fifteen units system. The experimental results are compared to other methods such as GA and PSO [7, 17]. The optimization goal for the fitness function as equation (13) is to minimize the outcome and then to dispatch ELD for power outputs. The parameters setting for Evolved Bat algorithm is the initial the total population size *n* = 20 and the dimension of the solution space *D* = 6 for six units system (*D*=15 for fifteen units system). Fitness function contains the full iterations of 200 is repeated by different random seeds with 10 runs. The final results are obtained by taking the average of the outcomes from all runs. The results are compared with the GA and PSO for ELD respectively.

The parameters setting for PSO (for further setting to reference in [18]) is the initial inertia weight $W=$ *(0.9−07*rand),* coefficients of learning factors c_1=-2 and c_2=2 in PSO, the population size n = 20 and the dimension of the solution space D = 6 for six units system (D=15 for fifteen units system). The parameters setting for GA (for further setting to reference in [19]) is the population size n = 20 and the dimension of the solution space D = 6 for six units system (D=15 for fifteen units system), probability crossover rate p_c=0.8, mutation rate p_m=0.01 in GA. Fitness function contains the full iterations of 200 is repeated by different random seeds with 10 runs.

5.1 Case Study 1- Six Units System

In this example, a system with six thermal units is used to demonstrate how the proposed approach works. The load demand is 1200MW. The characteristics of the six thermal units are given in Tables 1. In this case, each solution contains six generator power outputs, such as P1, P2, P3, P4, P5 and P6. Initialization of solution for power generating units is generated randomly. The dimension of the population is equal to 6. In normal operation of the system, the loss coefficients with the 100-MVA base capacity are as follows:

$$B_{ij} = 10^{-3} \times \begin{bmatrix} 0.14 & 0.17 & 0.15 & 0.19 & 0.26 & 0.22 \\ 0.17 & 0.60 & 0.13 & 0.16 & 0.15 & 0.20 \\ 0.15 & 0.13 & 0.65 & 0.17 & 0.24 & 0.19 \\ 0.19 & 0.16 & 0.17 & 0.71 & 0.30 & 0.25 \\ 0.26 & 0.15 & 0.24 & 0.30 & 0.69 & 0.32 \\ 0.22 & 0.20 & 0.19 & 0.25 & 0.32 & 0.85 \end{bmatrix} \tag{15}$$

$$B_0 = 10^{-3}[-0.390\ -0.129\ 0.714\ \ 0.059\ \ 0.216 - 0.663] \tag{16}$$

$$B_{00} = 0.056;\ \ P_D = 1200\text{MW} \tag{17}$$

Table 1. Generating unit's capacity with 1200MW load demand and coefficents

Unit	γ $/MW^2	β $/MW	α $	P$_{\min}$MW	Pmax MW
1	0.0070	7.0	240	100	500
2	0.0095	10.0	200	50	200
3	0.009	38.5	220	80	300
4	0.0090	11.0	200	50	150
5	0.0080	10.5	220	50	200
6	0.0075	12.0	120	50	120

Table 2 compares the statistic optimal results of three methods of GA, PSO and EBA for ELD in the system with six-generator of thermal plant. These three methods are run 10 times by different random seeds with full iterations of 200 for evaluating the fitness function of equation (13). That involved the average of 10 runs for generation power outputs, generation total cost, total power loss value, and total CPU times respectively. It is clearly seen that the method of EBA for ELD had total generating cost ($/h) is smallest with 14845 while its figures of this for GA and PSO are 14862 and 14861 respectively. The power loss of EBA method is also smaller than other methods as GA and PSO for ELD. In addition, the speed of running of EBA method is quite fast, it is faster than GA method for ELD.

Table 2. The best power output for six-generator system

Unit Output	GA	PSO	EBA
P1 (MW)	459.5422	458.0056	459.2242
P2 (MW)	166.6234	178.5096	171.5735
P3 (MW)	253.0350	257.3487	255.4936
P4 (MW)	117.4263	120.1495	119.8301
P5 (MW)	153.2482	143.7840	154.7214
P6 (MW)	85.88567	76.75549	73.76758
Total Power Output (MW)	**1235.7610**	**1234.5530**	**1234.5300**
Total Generation Cost ($/h)	**14862**	**14861**	**14845**
Power Loss (MW)	**35.7610**	**34.5530**	**34.5300**
Total CPU time (sec)	256	201	206

Figure 1 shows the three curves in distribution outline of the best solution for 200 iterations is repeated 10 trials of three methods GA, PSO and EBA for ELD with six-generating unit system. It clearly can be seen that the applied method of EBA for ELD (solid red line) is the more convergence in compare with GA and PSO for ELD in the same condition.

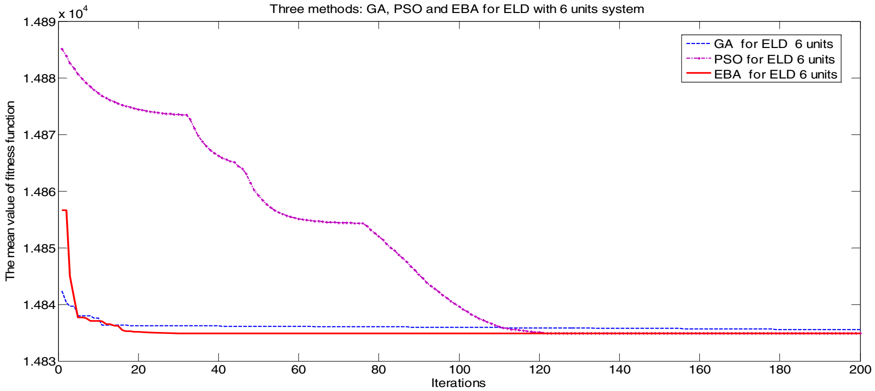


Fig. 1. Convergence characteristic of three methods GA, PSO, and the applied scheme of EBA for fifteen-generator system

5.2 Case Study 2- Fifteen Units System

In this example, a system contains fifteen thermal units for using to demonstrate how the proposed approach works. The load demand is 1700MW. The characteristics of the fifteen thermal units are given in Tables 3. Each solution contains fifteen generator power outputs, as listed from P1 to P15. The dimension of the population for this example is set to 15.

Table 3. Generating unit's capacity with 1700MW power demand and coefficents

Unit	γ \$/MW^2	β \$/MW	α \$	P_{min}MW	P_{max}MW
1	0.000230	10.5	670	150	455
2	0.000185	10.6	575	150	455
3	0.001125	8.5	375	20	130
4	0.001125	8.5	375	20	130
5	0.000205	10.5	460	150	470
6	0.000300	10.0	630	135	460
7	0.000362	09.7	549	135	465
8	0.000338	11.3	227	60	300
9	0.000807	11.2	173	25	162
10	0.001203	10.7	175	25	160
11	0.003587	10.3	187	20	80
12	0.005513	09.9	231	20	80
13	0.000371	13.1	225	25	85
14	0.001929	12.3	309	20	60
15	0.004547	12.4	325	15	55

In normal operation of the system, the loss coefficients with the 100-MVA base capacity are as follows,

$$B_{ij} = \quad 10^{-3} \times$$
$$\begin{bmatrix}
1.4 & 1.2 & 0.7 & -0.1 & -0.3 & -0.1 & -0.1 & -0.1 & -0.3 & -0.5 & -0.3 & -0.2 & 0.4 & 0.3 & -0.1 \\
1.2 & 1.5 & 1.3 & 0.0 & -0.5 & -0.2 & 0 & 0.1 & -0.2 & -0.4 & -0.4 & 0.0 & 0.4 & 1 & -0.2 \\
0.7 & 1.3 & -7.6 & -0.1 & -1.3 & -0.9 & -0.1 & 0.0 & -0.8 & -1.2 & -1.7 & 0.0 & -2.6 & 1.1 & -2.8 \\
-0.1 & 0.0 & -0.1 & 3.4 & -0.7 & -0.4 & 1.1 & 5.0 & 2.9 & 3.2 & -1.1 & 0.0 & 0.1 & 0.1 & -2.6 \\
-0.3 & -0.5 & -1.3 & -0.7 & 0.9 & 1.4 & -0.3 & -1.2 & -1 & -1.3 & 0.7 & -0.2 & -0.2 & -2.4 & -0.3 \\
-0.1 & -0.2 & -0.9 & -0.4 & 1.4 & 1.6 & 0 & -0.6 & -0.5 & -0.8 & 1.1 & -0.1 & -0.2 & -1.7 & 0.3 \\
-0.1 & 0.0 & -0.1 & 1.1 & -0.3 & 0.0 & 1.5 & 1.7 & 1.5 & 0.9 & -0.5 & 0.7 & 0.0 & -0.2 & -0.8 \\
-0.1 & 0.1 & 0.0 & 5.0 & -1.2 & -0.6 & 1.7 & 6.8 & 8.2 & 7.9 & -2.3 & -3.6 & 0.1 & 0.5 & -7.8 \\
-0.3 & -0.2 & -0.8 & -1.0 & -0.5 & 1.5 & 8.2 & 12.9 & -2.1 & -2.5 & 0.7 & -1.2 & 7.2 & 0.6 & -0.7 \\
-0.5 & -0.4 & -1.2 & 3.2 & -1.3 & -0.8 & 0.9 & 7.9 & 11.6 & 2.0 & -2.7 & -3.4 & 0.9 & -1.1 & -8.8 \\
-0.3 & -0.4 & -1.7 & -1.1 & 0.7 & 1.1 & -0.5 & -2.3 & -2.1 & -2.7 & 1.4 & 0.1 & 0.4 & -3.8 & 16.8 \\
-0.2 & 0.0 & 0.0 & 0.0 & -0.2 & -0.1 & 0.7 & -3.6 & -2.5 & -3.4 & 0.1 & 5.4 & -0.1 & -0.4 & 2.8 \\
0.4 & 0.4 & -2.6 & 0.1 & -0.2 & -0.2 & 0.0 & 0.1 & 0.7 & 0.9 & 0.4 & -0.1 & 10.3 & -10.1 & 2.8 \\
0.3 & 0.1 & 11.1 & 0.1 & -2.4 & -1.7 & -0.2 & 0.5 & -1.2 & -1.1 & -3.8 & -0.4 & -10.1 & 5.7 & -9.4 \\
-0.1 & -0.2 & -2.8 & -2.6 & -0.3 & 0.3 & -0.8 & -7.8 & -7.2 & -8.8 & 1.6 & 2.8 & 2.8 & -9.4 & 1.3
\end{bmatrix} \tag{18}$$

$$B_{i0} = 10^{-3} \times [-0.1 - 0.2\ \ 2.8 - 0.1\ 0.1 - 0.3 - 0.2 - 0.2\ 0.6\ 3.9 - 1.7\ 0.0 - 3.2\ 6.7 - 6.4] \tag{19}$$

$$B_{00} = 0.0055,\ P_D = 1700\ \text{MW}. \tag{20}$$

Table 4 provides statistic optimal results of three methods of GA, PSO and EBA for ELD in the system with fifteen-generator of thermal plant. The statistic results that are involved the generation power outputs, generation total costs, evaluation values, and average CPU time respectively. These three methods are run 10 times by different random seeds with full iterations of 200 for evaluating the fitness function of equation (15). It clearly can be seen that the applied method of EBA for ELD had total generation cost, power loss and the speed of EBA method are smaller than GA and PSO method for ELD.

Table 4. the best power output for fifteen-generator system

Unit Output	**GA**	**PSO**	**EBA**
P1 (MW)	455.00	455.00	455.00
P2 (MW)	93.99	123.03	80.00
P3 (MW)	82.06	58.85	100.93
P4 (MW)	89.97	75.56	45.29
P5 (MW)	150.00	162.94	150.00
P6 (MW)	350.76	322.48	357.49
P7 (MW)	226.36	165.70	242.22
P8 (MW)	60.00	60.34	60.56
P9 (MW)	52.37	91.84	27.60
P10(MW)	25.10	45.10	50.40
P11(MW)	25.96	42.70	30.60
P12(MW)	74.01	77.97	80.00
P13(MW)	66.99	45.38	66.27
P14 (MW)	34.22	47.37	26.24
P15 (MW)	51.05	55.00	55.00
Total Power Output (MW)	**1827.84**	**1837.27**	**1827.60**
Total Generation Cost ($/h)	**1534.66**	**1535.06**	**1534.61**
Power Loss (MW)	**137.84**	**129.27**	**127.60**
Total CPU time (sec)	376	303	302

Figure 2 shows the comparison of the applied method of EBA for ELD with 15 generating units system in convergence property distribution outline of the best solution for 200 iterations is repeated 10 trials, with GA and PSO for ELD in the same condition. It clearly can be seen that the applied method of EBA for ELD (solid red line) is the more convergence in compare with GA and PSO for ELD in the same condition.

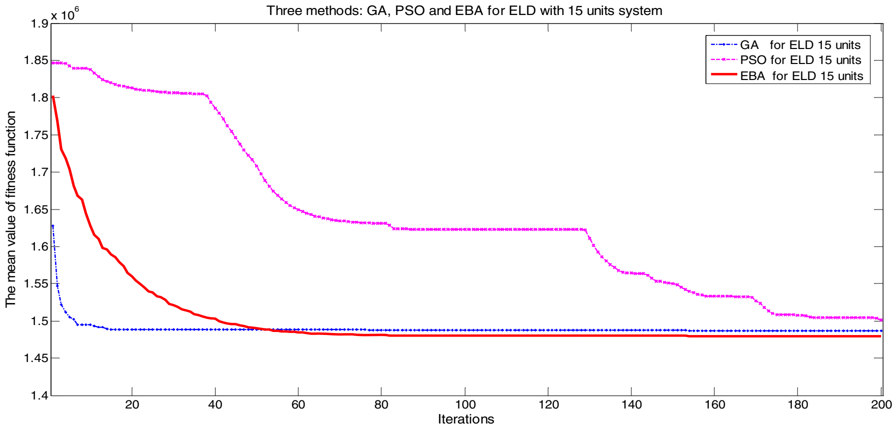


Fig. 2. Convergence characteristic of three methods GA, PSO, and the applied scheme of EBA for fifteen-generator system

From the experimental results in table 2, figure 1, table 4 and figure 2, it is clear that the applied method of EBA for ELD problem has the ability to find the better quality solution and has better convergence characteristics, computational efficiency and less average CPU time when compared to other methods such as GA and PSO for ELD problem

6 Conclusion

In this paper, an application of the Evolved Bat algorithm for solving ELD problem was presented. To evaluate the solution quality and computation efficiency of this application of the Evolved Bat algorithm for ELD problem, the case study with systems had six units and fifteen units of thermal plants are implemented to solve ELD problem with the generator constraints as linear equality and inequality constraints and also considering transmission loss. The implemented results were compared with the genetic algorithm (GA) method for ELD, and with the particle swarm optimization (PSO) method for ELD, show that the applied EBA method for ELD can provide the better quality solution, the higher efficiency and accuracy, and less average CPU time.

References

1. Allen, B.F.W., Wood, J.: Power Generation, Operation and Control, 3rd edn., p. 656. Wiley (2013)
2. Han, X.S., Gooi, H.B.: Effective economic dispatch model and algorithm. International Journal of Electrical Power & Energy Systems 29(2), 113–120 (2007)

3. Dillon, T.S., Edwin, K.W., Kochs, H.D., Taud, R.J.: Integer Programming Approach to the Problem of Optimal Unit Commitment with Probabilistic Reserve Determination. IEEE Transactions on Power Apparatus and Systems PAS-97(6), 2154–2166 (1978)
4. Liang, Z.X., Glover, J.D.: A zoom feature for a dynamic programming solution to economic dispatch including transmission losses. IEEE Transactions on Power Systems 7(2), 544–550 (1992)
5. Jong-Bae, P., Ki-Song, L., Joong-Rin, S., Lee, K.Y.: A particle swarm optimization for economic dispatch with nonsmooth cost functions. IEEE Transactions on Power Systems 20(1), 34–42 (2005)
6. Zwe-Lee, G.: Particle swarm optimization to solving the economic dispatch considering the generator constraints. IEEE Transactions on Power Systems 18(3), 1187–1195 (2003)
7. Po-Hung, C., Hong-Chan, C.: Large-scale economic dispatch by genetic algorithm. IEEE Transactions on Power Systems 10(4), 1919–1926 (1995)
8. Bakirtzis, A., Petridis, V., Kazarlis, S.: Genetic algorithm solution to the economic dispatch problem. IEE Proceedings Generation, Transmission and Distribution 141(4), 377–382 (1994)
9. Wong, K.P., Wong, Y.W.: Genetic and genetic/simulated-annealing approaches to economic dispatch. IEE Proceedings Generation, Transmission and Distribution 141(5), 507–513 (1994)
10. Simopoulos, D.N., Kavatza, S.D., Vournas, C.D.: Unit commitment by an enhanced simulated annealing algorithm. IEEE Transactions on Power Systems 21(1), 68–76 (2006)
11. Chowdhury, B.H., Rahman, S.: A review of recent advances in economic dispatch. IEEE Transactions on Power Systems 5(4), 1248–1259 (1990)
12. Tsai, J.-S.P.P.-W., Liao, B.-Y., Tsai, M.-J.: Bat Algorithm Inspired Algorithm for Solving Numerical Optimization Problems. Applied Mechanics and Materials 148-149(2012), 134–137 (2011)
13. Yang, X.-S.: A New Metaheuristic Bat-Inspired Algorithm. In: González, J.R., Pelta, D.A., Cruz, C., Terrazas, G., Krasnogor, N. (eds.) NICSO 2010. SCI, vol. 284, pp. 65–74. Springer, Heidelberg (2010)
14. [14] Rao, S.S., Rao, S.S.: Engineering Optimization: Theory and Practice. Wiley (2009)
15. Arora, J.S.: Introduction to Optimum Design. Academic Press (2011)
16. Michalewicz, Z., Schoenauer, M.: Evolutionary algorithms for constrained parameter optimization problems. Evol. Comput. 4(1), 1–32 (1996)
17. Zwe-Lee, G.: “Closure to “Discussion of ’Particle swarm optimization to solving the economic dispatch considering the generator constraints”. IEEE Transactions on Power Systems 19(4), 2122–2123 (2004)
18. Yuhui, S., Eberhart, R.C.: Empirical study of particle swarm optimization, vol. 3, p. 1950 (1999)
19. Srinivas, M., Patnaik, L.M.: Genetic algorithms: a survey. Computer 27(6), 17–26 (1994)

A Simple and Accurate Global Optimizer for Continuous Spaces Optimization

Zhenyu Meng* and Jeng-Shyang Pan

Innovative Information Industry Center,
Computer Science Department,
Harbin Institute of Technology Shenzhen Graduate School,
Shenzhen, China
{mzy1314,jengshyangpan}@gmail.com

Abstract. Ebb-Tide-Fish Algorithm (ETFA) is a simple but powerful optimization algorithm over continuous search spaces, and the inspiration comes from the foraging behavior of the fish in ebb tide. This kind of fish is a fascinating creature, and it often draws my attention when I walk on the beach. When I studied and got an idea of improving some optimization algorithms recently, the kind of fish flashes in my mind. The algorithm mainly focuses on the diversity of locations of the fish rather than what velocity it is when the fish swim from the current location to a better one. The algorithm gives a formulation of the foraging behavior of the fish, and the detailed model is also given in the paper. The performance of ETFA on a testbed of four functions is compared with several famous published methods. The final results show that ETFA has a faster convergence rate with an excellent accuracy.

Keywords: Benchmark function, Ebb-Tide-Fish Algorithm, Optimization, Particle Swarm Optimization.

1 Introduction

There are many kinds of demands for tough optimization problems in our lives. The standard approach to tackle these problems often begins by designing an objective function that can model the problems' objectives while incorporating any constraints [1]. Optimization problems, such as hardware design, circuit design, electricity supply, parameter tuning, time scheduling and so on, often have different object functions, even some of them do not need a specific object function but operate with so-called regions of acceptability. Obviously, the optimization problem do not need an objective function are usually inferior to techniques with one. Therefore, we focus mainly on those with an objective function. We also use a lot of benchmark function/objective function to validate our proposed algorithm.

* Corresponding author.

H. Sun et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 329, DOI: 10.1007/978-3-319-12286-1_13

As we know, what is sought is a general-purpose algorithm that can optimize an arbitrary function without user intervention [1, 5]. So many evolutionary algorithms have been proposed as this goal is unable to completely achieved. PSO, Simulated Annealing, Genetic Algorithm, Differential Evolution, Ant Colony Optimization and Harmony Search have been proposed for ages while there also are some algorithms proposed in the recent years, such as Cat Swarm Optimization (CSO), Cuckoo Search Algorithm, Firefly Algorithm (FA), Bat Algorithm (BA) etc [2–4]. They all are powerful optimization methods for specific tough problems and also have disadvantages in their own.

Like some other algorithms, ETFA simplifies the optimization problem into two steps, intensification and diversification. That one fish swims toward temporary optima of the population completes the diversification of solutions, and the local search completes intensification. Several algorithms use both velocity and locations to complete diversification, such as PSO, CSO, FA, BA and so on. In the ETFA algorithm, we mainly focus on the location information and the experiment results show that the proposed algorithm outperforms the ones that use both velocity and location for optimization.

ETFA's update scheme is very easy and simple to implement. There are two main approach, one can be describe as the movement toward current global optima, and the other is local searching. The proposed algorithm only uses two parameters for optimization. One is random ratio that defines the moving distance, and the other is the local searching scale. However, the population size must be decided at run-time for the optimization problem.

Once a variation is generated, the algorithm gives a strategy to decide whether to accept the newly derived parameters or not. Under the common greedy criterion, the reduced value of the benchmark function is accepted to renew the old parameters. After a fixed number of iteration or exist some other criteria, the algorithm ends with the final global optima value outputted in the console window.

We formulate the new revolution algorithm, use it to tackle some tough optimization problems, and also give the algorithm a name, called Ebb-Tide-Fish algorithm. The rest of the paper is organized as follows. Firstly, we will describe the algorithm in section 2. Section 3 and section 4 give the benchmark functions and simulation results respectively. Then we will give a conclusion and future work in Section 5.

2 The Ebb-Tide-Fish Algorithm

Many species of animals or insects show some intelligent social behaviors. Some species have a lead in their group while in others each individual has self-organized behaviors. These enable them to communicate with each other and make perceptions of the living circumstance. Some can learn knowledge about/from their group and environment while some only have inborn ability for survival.

The behaviors of more intelligent creatures are too complex to establish a model to solve specific problems. However, for some less intelligent creatures, the inborn ability can be drawn as good analogy to produce products or to propose new formulated algorithms. Fish is a good example of swarm intelligence. There are also some other similar algorithms such as bacteria algorithm, particle swarm optimization, article bee optimization etc.

The kind of fish that foraging and swimming in the ebb tide is very fascinating. It drew my attention when I walked on the beach, and the behavior of this kind of fish flashed in my mind when I wanted to improve the performance of some optimization algorithm. Now we propose the algorithm to formulate the intelligent fish to tackle some optimization problems. When we had a survey of different kinds of optimization algorithms, we found that [8] had proposed an artificial fish swarm algorithm. The artificial fish swarm algorithm is a powerful fish-inspired algorithm, but our proposed algorithm is a total different one, which is much simpler, and of better convergence rate and accuracy.

Now we will give a detailed description of parameters and formulas used in the algorithm. The parameter F_j denotes the j^{th} fish in a population. X_j denotes the coordinates of the j^{th} fish. There are n populations of fishes and each fish in a D-dimension search domain, and these can be represented in equation 1 and 2 respectively. The benchmark function is given in $f(X_j)$ form.

$$F = F_1, F_2, ..., F_n \tag{1}$$

$$X_j = (x_1, x_2, ..., x_d) \tag{2}$$

There are some fish in the population are labeled as single search ones, while others swim following the population to the global optima. For the single search approach, the fish would also like to search a nearby place in diversified ways. So it will swim about for N times ($N = \frac{d_{count}}{2} * \beta$, and d_{count} denotes the number of dimensions in the searching domain. For simplification we use a fixed number $\beta = 5.$). As we know, only a small proportion of fish would like to search individually, others would like to forage following the population. So we set the proportion of the population for single swim search is $rate = 0.01$. Figure 1 shows the search behavior of fish. Let $X_j = (x_1, x_2, ..., x_d)$ denotes the coordinate of the j^{th} fish, $X_{center} = (x_{1_0}, x_{2_0}, ..., x_{d_0})$ denotes the searching domain center. The local searching radius of single search fish is $r = \frac{1}{\beta} * (x_i - x_{i_0})$ in each dimension. For the searching toward global optima approach, they swim a random distance from the current location direct to the temporary global optima. The pseudo code of the algorithm is described in the Algorithm 1.

3 Benchmark Function for Optimization

For the optimization problem, users generally demand the related algorithm to fulfill several requirements, such as ability to handle nonlinear, multi-modal functions, parallelizability to deal with intensive computation, convergence, robustness etc. Many benchmark functions are also proposed to validate the optimization algorithms. These functions, known as artificial landscapes in applied mathematics, are used to evaluate characteristic of the proposed algorithms, such as:

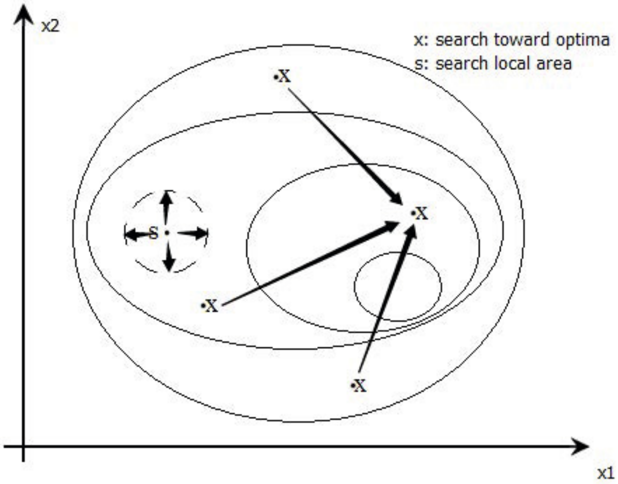


Fig. 1. The search behavior of the fish

- Velocity of convergence
- Precision
- Robustness
- General performance

Velocity of convergence is important for computational demanding optimizations. It describes the time consumed till the algorithm gets the optima value. This may varies from several second to several days. A good algorithm saves lots of treasure and time. Precision is also an important characteristic of the optimization, as most algorithms do not grantee global optima. Although simulated annealing acclaims to get global optima, it is only achieved when the cooling process is slow enough and the simulation time is long enough. Some algorithms often fall in local optima rather than finding the global one. Robustness characteristic means that the algorithm can run well although there exist some noises. The last but not the least characteristic is general performance, which gives a general evaluation of the algorithm.

In this paper, we use several test functions, the first one is the Rosenbrock function, it is a non-convex function and introduced by Howard H. Rosenbrock in 1960. Equation 3 shows the formula of the function. The global minimum is inside a long, narrow, parabolic shaped flat valley, and to find it is trivial and so it is difficult. For a D-dimension function, the minimum value is achieved when the coordinate of each dimension is 1 ($f(1, 1, ..., 1) = 0$). Figure 2 shows the mesh plot of the function.

$$f(X(x_1, x_2, ..., x_d)) = \sum_{i=1}^{d-1} (100(x_{i+1} - x_i^2)^2 + (x_i - 1)^2) \tag{3}$$

Algorithm 1. Pseudo code of the algorithm

```
Initialization:
    Initialize the searching space $V$, locations of fish $X$, and the benchmark function $f(X)$.
Iteration:
while $exeTime < MaxIteration || !stopCriteria$ do
    if $localSearch = True$ then
        Search around local scope by jumping to the bound and find the local best $X_{lbest}$.
        $X_{pos_i} \leftarrow X_{lbest}^t$.
        Reset local search flag.
    else
        Search toward global optima.
        $X_{pos_i} \leftarrow X_i^t + (X_{gbest}^t - X_i^t) * rand()$.
    end if
    Set local search fish flags.
    Calculate the fitness value of the current locations and label the current global best $X_{gbest}$.
end while

Output:
    The global optima $X_{gbest}$.
```

The second test function we use in the paper is Sphere function in Equation 4. The global minimum is $f(0,0,...,0) = 0$. The last one is Beale's function with the formula in Equation 5. The global minimum is $f(3, 0.5) = 0$. Figure 3 shows the distribution of fish population in Sphere function after the 4^{th} iteration.

$$f(X(x_1, x_2, ..., x_d)) = \sum_{i=1}^{d} x_i^2 \tag{4}$$

$$f(x, y) = (1.5 - x + x * y)^2 + (2.25 - x + x * y^2)^2 + (2.625 - x + x * y^3)^2 \tag{5}$$

4 Simulation Results of the Proposed Algorithm

In order to validate and evaluate the proposed algorithm, comparisons are made to make a glance of the performance of the algorithm. PSO, Bat Algorithm, Different Evolution, Cat Swarm Optimization and the proposed algorithm are compared here. The parameters of the population size in each algorithms are the same for listed algorithms above. There are two main approach for evaluation, one is to compare the accuracies for a fixed number of function evaluations, and the other is to compare the numbers of function evaluations for a given tolerance or accuracy. We use the first approach for evaluation.

There are many derivation versions of PSO, for simplification, we use the standard version of it, with learning parameters $c1 = c2 = 2$ in the Equation 6. For Bat Algorithm, the default parameters do not have good convergence

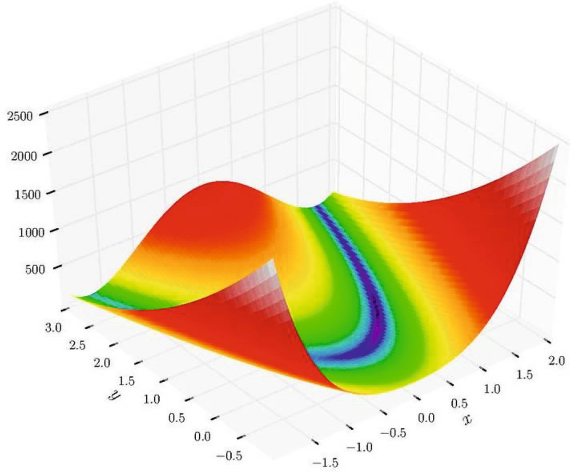


Fig. 2. The Rosenbrock Function

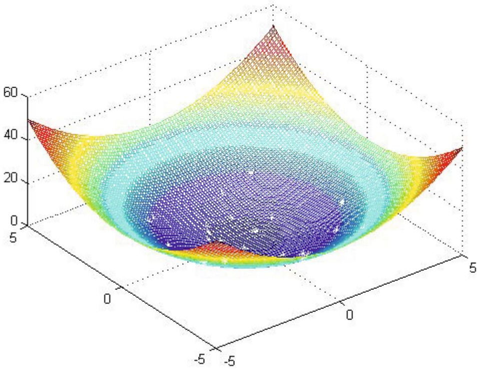


Fig. 3. Fish distribution after 4th iteration in sphere function

rate in our experiment, and the author use the second approach (comparing the numbers of function evaluations for a given accuracy) for comparison. But the parameters for BA in PSO form (BA(pso)) and Harmony search from (BA(hs)) have a good convergence rate. So we use the above two form as comparisons of BA. For CSO, we use the default value of the number of seeking pool ($smp = 5$). For the DE, we use the "DE/best/1/bin" form to make comparison. Figure 4 shows the convergency rate of DE under a fixed number of function evaluation (150). We run ten times and only three times can reach the accuracy under the value 0.05 of benchmark function Rosenbrock. The performance of the proposed algorithm (ETFA) for the Rosenbrock benchmark function can be seen in Figure 5. Figure 6 shows the comparison for Sphere benchmark function.

$$v_i = v_i + c1 * (p_{i_{pre}} - x_i) + c2 * (p_{gbest} - x_i) \tag{6}$$

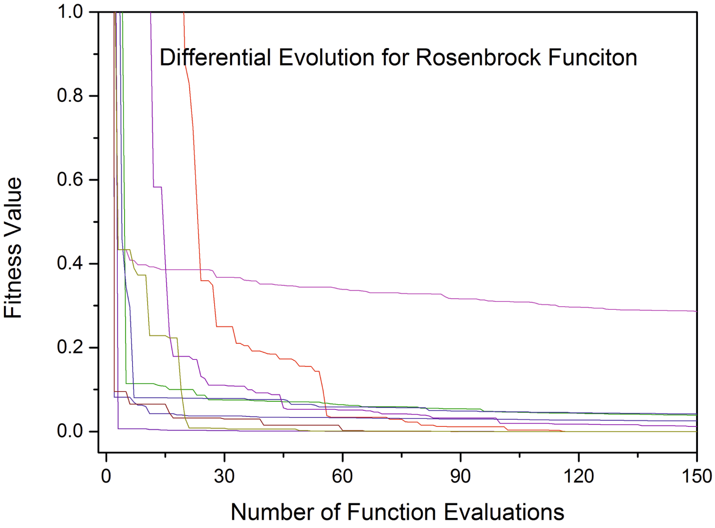


Fig. 4. Accuracy in fixed number of function evaluation for differential evolution of Rosenbrock function

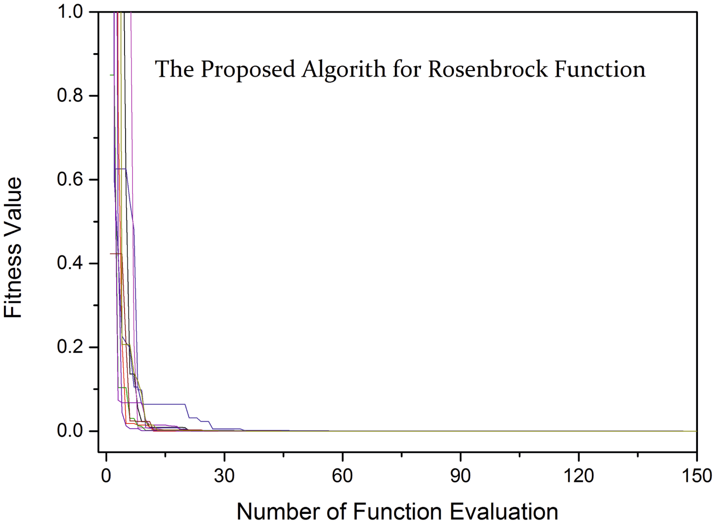


Fig. 5. Accuracy in fixed number of function evaluation for ETFA of Rosenbrock function

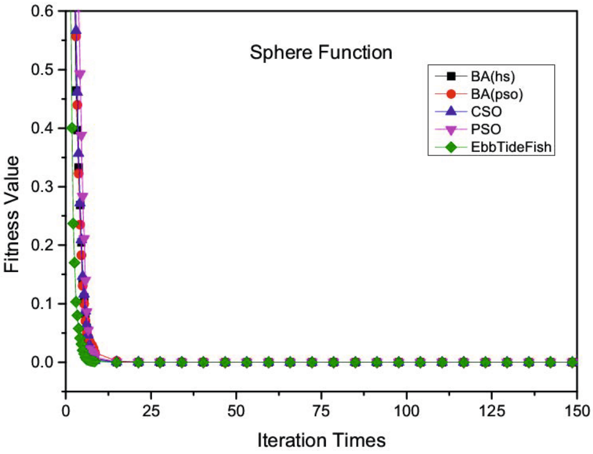


Fig. 6. Comparisons of different algorithms under Sphere benchmark function

5 Conclusion and Future Work

The Ebb-Tide-Fish Algorithm (ETFA) for optimization has been introduced and compared to Particle Swarm Optimization (PSO), Bat Algorithm (BA), Cat Swarm Optimization (CSO), and Differential Evolution (DE). The performance of ETFA is very well and outperforms some of the optimization algorithms in terms of required number of function evaluations with tolerant accuracy, although there still are spaces for improvement. It is simple and straightforward and is still in its infancy. The time complexity of the proposed algorithm is still excellent than some other algorithms. A future work is to make some improvements and make a clear sense of the scaling property and behavior of the algorithm.

References

1. Storn, R., Price, K.: Differential evolutionCa simple and efficient heuristic for global optimization over continuous spaces. Journal of Global Optimization 11(4), 341–359 (1997)
2. Kennedy, J.: Particle swarm optimization. In: Encyclopedia of Machine Learning, pp. 760–766. Springer, US (2010)
3. Yang, X.-S.: A new metaheuristic bat-inspired algorithm. In: González, J.R., Pelta, D.A., Cruz, C., Terrazas, G., Krasnogor, N. (eds.) NICSO 2010. SCI, vol. 284, pp. 65–74. Springer, Heidelberg (2010)
4. Puranik, P., Bajaj, P., Abraham, A., Palsodkar, P., Deshmukh, A.: Human Perception-based Color Image Segmentation Using Comprehensive Learning Particle Swarm Optimization. Journal of Information Hiding and Multimedia Signal Processing 2(2), 227–235 (2011)

5. Clerc, M., Kennedy, J.: The particle swarm-explosion, stability, and convergence in a multidimensional complex space. IEEE Transactions on Evolutionary Computation 6(1), 58–73 (2002)
6. Engelbrecht, A.P.: Fundamentals of computational swarm intelligence. John Wiley & Sons (2006)
7. Chu, S.-C., Tsai, P.-w., Pan, J.-S.: Cat swarm optimization. In: Yang, Q., Webb, G. (eds.) PRICAI 2006. LNCS (LNAI), vol. 4099, pp. 854–858. Springer, Heidelberg (2006)
8. Neshat, M., Sepidnam, G., Sargolzaei, M., et al.: Artificial fish swarm algorithm: a survey of the state-of-the-art, hybridization, combinatorial and indicative applications. Artificial Intelligence Review, 1–33 (2012)

Spatial Evolutionary Algorithm for Large-Scale Groundwater Management

Jihua Wang[1], Ximing Cai[2], and Albert Valocchi[2]

[1] Department of Forecasting, FedEx Corporate Services, 3640 Hacks Cross Rd, Memphis, TN, USA
[2] Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, USA
jwang41@gmail.com, {xmcai,valocchi}@illinois.edu

Abstract. Large-scale groundwater management problems pose great computational challenges for decision making because of the spatial complexity and heterogeneity. This study describes a modeling framework to solve large-scale groundwater management problems using a newly developed spatial evolutionary algorithm (SEA). This method incorporates spatial patterns of the hydrological conditions to facilitate the optimal search of spatial decision variables. The SEA employs a hierarchical tree structure to represent spatial variables in a more efficient way than the data structure used by a regular EA. Furthermore, special crossover, mutation and selection operators are designed in accordance with the tree representation. In this paper, the SEA was applied to searching for the maximum vegetation coverage associated with a distributed groundwater system in an arid region. Computational experiments demonstrate the efficiency of SEA for large-scale spatial optimization problems. The extension of this algorithm for other water resources management problems.

Keywords: Spatial Optimization, Genetic Algorithms, Large-Scale, Decision making, Resources allocation.

1 Introduction

In general spatial optimization is a methodology used to optimize a management objective by searching an appropriate pattern of certain spatial variables, given finite resources, and spatial relationships in an environmental system[1]. The methodology is challenging because the spatial patterns are usually implicit and it is difficult to represent them in quantitative models. On the other hand, the methodology is promising because the knowledge of the spatial patterns is informative for model design [2] and useful to solve large-scale, computationally expensive models, particularly optimization models. This paper presents an optimization methodology that incorporates the knowledge of spatial patterns with the design of evolutional algorithm and applies the algorithm to solving a large-scale ecological restoration which contains more than 10,000 decisions variables.

H. Sun et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 329, DOI: 10.1007/978-3-319-12286-1_14

Evolutionary algorithms (EA) have been demonstrated to be successful in solving optimization models for water resources management due to their flexibility in incorporating complex simulation models in optimal search procedures [3-5]. However, a regular EA (REA) has limited capability in solving large-scale optimization models. In particular, groundwater management problems that this study focuses on involve two-dimensional (2-D) variables. SEA modifies the encoding and operators of EA, and assimilates spatial information to make it more computationally effective for spatial problems than REA [6].

Idea of using spatial information to enhance EA for complex spatial optimization models was systematically discussed by Openshaw [7-8]. Since then researchers from geography and computer science have developed SEA for site-search [9-10], image segmentation [11], K-means clustering [12], domain decomposition in spatial interpolation [13], etc. The key procedure is to incorporate spatial information for encoding schemes and modifying EA operators such as mutation and crossover. A special data structure is required for the realization of the procedure. Xiao et al. [14] and Brooks et al. [8] employed graph as a new encoding schemes to represent the EA solutions and modified the EA operators to maintain spatial contiguity. More recently, Cao et al [15] used grid to represent a land use solution and developed a non-dominated sorting genetic algorithm-II for multi-objective optimization of land use (NSGA-II-MOLU) to search for optimal land use scenarios with multiple objectives and constraints reflecting the requirements of land users. Fotakis and Sidiropoulos [16] developed a multi-objective self-organizing algorithm (MOSOA) based on cellular automata for a combined land use planning and resource allocation problem. However, none of these studies has tested their method for large-scale optimization problems with more than 500 decision variables. Gong and Yang [11] and Laszlo and Mukherjee [17] used a tree-based EA for image processing and showed the effectiveness of the algorithm for image processing problems. . In this paper, we also use the tree as an encoding scheme with a hierarchical structure to represent the solutions of groundwater management problems, and tailor the algorithm development to the spatial specialties of the studying problems and employ the spatial specialties to re-design the EA operators. We demonstrate the procedures of SEA through a testing problem.

2 Spatial Evolutionary Algorithm (SEA)

Compared to a regular EA, the essential procedure of SEA is to utilize the spatial information in the algorithm design and further clarify spatial patterns associated with the modeling problem. Using a top-down method, SEA starts from an initial spatial pattern of a decision variable and then further refines the pattern as evolution proceeds. The accuracy of the refinement depends on the needs of the decision makers. This method can then balance the tradeoff between accuracy and computation and hence provides flexibility for solving practical problems.

In this study, the SEA employs a hierarchical tree structure to represent spatial variables. Instead of representing individual grids used in the physical simulation model, the tree structure represents a sub-set of grids by branch and leave. Furthermore, special crossover, mutation and selection operators are re-designed to incorporate the spatial information.

2.1 Tree-Based Data Structure

Encoding schemes for decision variables are fundamental to all EAs. In REA, solutions are encoded with binary strings [18], real numbers [19] or finite state machines [20]. Generally, the encoding method of the solutions depends not only on the context of the problem but also on the genetic operators used.

Xiao et al. [14] designed a graph encoding and corresponding EA operators to solve a multiobjective site-search problem. The spatial contiguity of a site must be maintained [14, 9, 21] for the site-search problems and hence an undirected graph is used to represent a contiguous solution. In their context, a space can be discretized into raster cells and each vertex in the graph represents a cell in the space and the four edges of this vertex represent the connections between this cell with its adjacent four cells. With this encoding scheme and the corresponding EA operators, the solution contiguity will persist through all generations of EA. Spatial contiguity is important for reserve network design and site search because a contiguous landscape provides physical condition and increases the opportunities for species dispersal and migration. However, many spatial optimization problems, including those in groundwater management, the contiguity of spatial variables is not a concern. The variables can be spatially distributed without explicit connections, for example, the vegetation coverage density patches fed by groundwater extraction [22] and groundwater pumping wells in different areas [23]. Moreover, the method in Xiao [14] focuses on the location and reconfiguration of a small number of patches (10 patches) and exerts more computational efforts on changing the location and shape of the site (e.g., identifying the neighborhood). However, the optimization of vegetation coverage associated with water table or the optimization of groundwater pumping by a large number of well in a large area has a much bigger search space of reconfiguration and the computational efforts will be spent mostly on the interesting subsets of the entire spatial domain.

Cao et al [15] used a list or grid of genes to represent a land use solution where the position of each gene (cell) represents a unit and the land use of the unit is determined by its value. They developed a NSGA-II for multi-objective optimization of land use (NSGA-II-MOLU). They applied NSGA-II-MOLU to search for optimal land use scenarios with multiple objectives and constraints extracted from the requirements of users. Although this method is efficient to searching over tens of thousands of solutions of trade-off sets for a multi-objective spatial optimization problem it must pre-determine the land use parcels at the very beginning and then optimize the solutions. However, the blocks or zonations used in SEA can evolve along the generations depending on the heterogeneity of the system and the computational facilities, which offers more flexibilities to the decision makers and modelers. Fotakis and Sidiropoulos [16] developed a multi-objective self-organizing algorithm (MOSOA) based on cellular automata to handle both local and global spatial constraints. This method is applied for a combined land use planning and resource allocation problems. The study area is divided into land blocks and each block includes a number of pumping wells in fixed positions. After optimization of land blocks, each block is assigned the land use type and the water sources. However, these blocks are fixed at the very beginning and the land use type is assigned uniformly in that block.

Brooks combined EA with a region-growing programme (RGP) for a site-search problem [10]. In his approach, several seed grids are selected firstly and then grown into sites with specified size. This algorithm exerts most computational efforts on controlling the growth orientation and shape of the sites, in order to get a contiguous site. This algorithm is limited with pre-determined seed grids and applied to problems of growing regions from several identified sites.

To solve large-scale spatial optimization problems, this study attempts to overcome some limitations discussed above by allowing the zonation (boundary) and the content within each zone to be improved simultaneously along the SEA generations.

Two essential features are needed for the design of the encoding scheme. First, since large-scale problems are computationally expensive if each grid of a map is encoded as a decision variable, it is necessary for the encoding of the population to represent the spatial features with limited data volume. Second, the spatial solutions must be represented by a well-defined spatial data structure that facilitates EA operators to adopt the spatial features. To meet these requirements, a tree structure is employed to represent the spatial solution because it has been shown to capture spatial features with reduced data volume [24]. In addition, as shown below, the operators of crossover and mutation can be designed to accommodate on the tree structure.

As an illustration example, a vegetation map is encoded as a quadtree to represent one individual in a SEA population. Every leaf without a predecessor in the quadtree (e.g., part A in Figure 1b) represents a uniform coverage in the population map (the blue square in Figure 1a) while a node (e.g., B in Figure 1b) represents four different vegetation patches (orange square in Figure 1a).

The tree structure of SEA offers flexibility for modelers to get various degrees of accuracy of the optimization solution for a large-scale spatial problem. The quadtree starts from a rough pattern and then further refines the map. The quadtree in Figure 1b, for example, starts from a uniform map and is constructed from level 1, as node C shows. If the refinement of the map leads to an improvement of the optimization objective, node C splits, and the depth of quadtree goes to level 2 and leads to more variations in the corresponding map. If we want to increase the accuracy further, node B in level 2 splits again, and the depth of quadtree goes to level 3, as shown in the map (Figure 1a). Refinement stops when further expansion of the tree does not result in additional fitness improvement as specified by a threshold.

It is worth noting that genetic programming (GP) [25] also uses a tree as the encoding scheme, but some important differences exist between these GP and SEA.

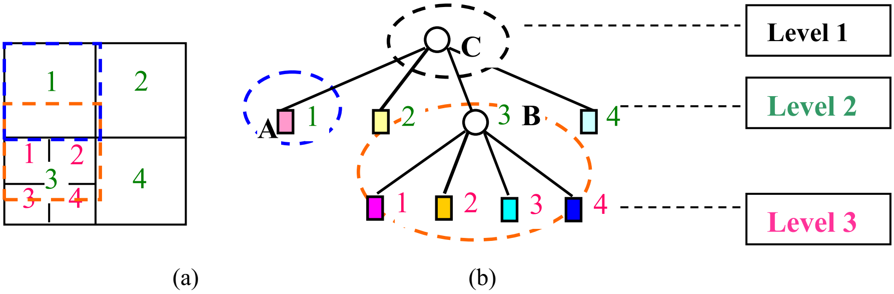


Fig. 1. Encoding a spatial individual (a, representing a vegetation density map) with a quadtree (b)

First, in SEA the nodes of the tree represent solutions at different spatial resolutions while the nodes in GP represent function elements that can be selected for model development. Second, the height of the tree in SEA is limited by the resolution levels as discussed in section 3.2 while the tree height in GP is unlimited.

2.2 Flowchart of SEA

As shown in Figure 2, SEA starts from initialization of population and then moves to crossover, mutation, selection and terminates when the stop condition is satisfied. The enhanced components have been marked in gray and will be discussed with more details later. The left loop of the flowchart checks the feasibility of the solution (e.g., using the groundwater simulation model to determine whether groundwater can support a projected vegetation coverage. If not, a new solution of vegetation coverage will be generated for further feasibility evaluation. The right loop represents the generations of SEA. The evolution of SEA will not stop until (1) generation reaches the maximum number of generations, (2) there is no improvement within a specified number of generations, and (3) a specified percentage of the population (popperc in Table 2) reaches the maximum height of the tree structure. The third stop condition is designed specifically for the hierarchical tree structure of SEA. The maximum height of the tree structure is determined by the resolution of the simulation model. For example, if the spatial domain is represented by N*N grids in a groundwater model that simulates water table corresponding to a specified vegetation density map, the maximum height is $\log_2(N)+1$, N can be 4, 8, 16,…128 and so on.

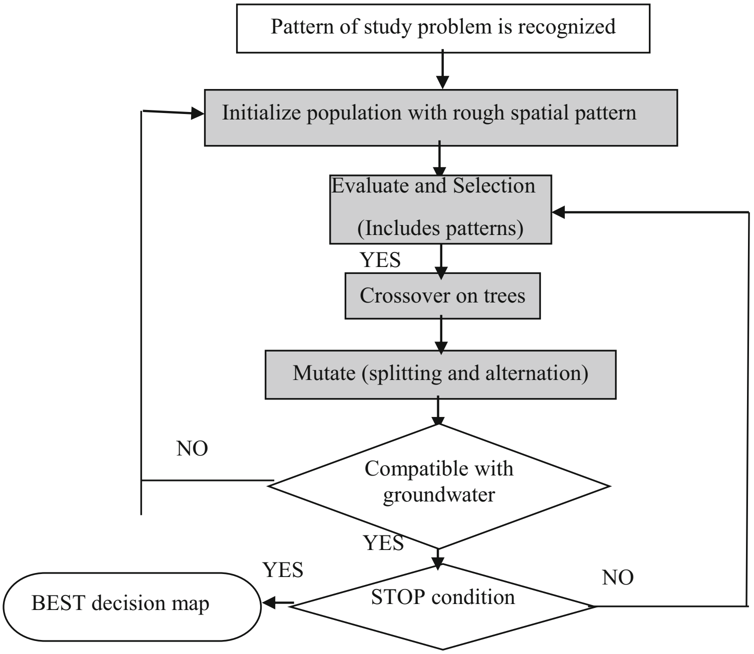

Pattern of study problem is recognized
Initialize population with rough spatial pattern
Evaluate and Selection
(Includes patterns)
YES
Crossover on trees
Mutate (splitting and alternation)
NO
Compatible with groundwater
YES
YES
BEST decision map
STOP condition
NO

Fig. 2. Flowchart of SEA, The left loop checks whether groundwater can support this vegetation coverage. The right loop represents the generations of SEA. The shadowed rectangles highlight the difference in SEA operators.

Table 1. Parameter settings of REA and SEA in the test case. Notes: popperc is the specified percentage of the population; swapperc is the specified percentage of swapping; senp is the probability of splitting sensitive leaves in mutation; rsplitp is the probability of randomly splitting in mutation; alterp is the probability for alternating leaves in mutation.

Parameter	REA	SEA
Population size	80	80
popperc		0.8
Crossover Probability (Pc)	0.8	0.8
swapperc		0.5
Mutation Probability (Pm)	0.08	0.5
senp		0.5
rsplitp		0.1
alterp		0.3
Encoding	Real	Quadtree
Crossover method	Arithmetic	Swapping trees
Mutation method	Non-uniform	Splitting and alteration
Selection method	Tournament	Includes patterns

2.3 Crossover on Trees

To encode the solution with a quadtree, the crossover and mutation operators must be modified to meet the spatial property of the tree structure and ensure that the results from the two operations are still legitimate quadtrees.

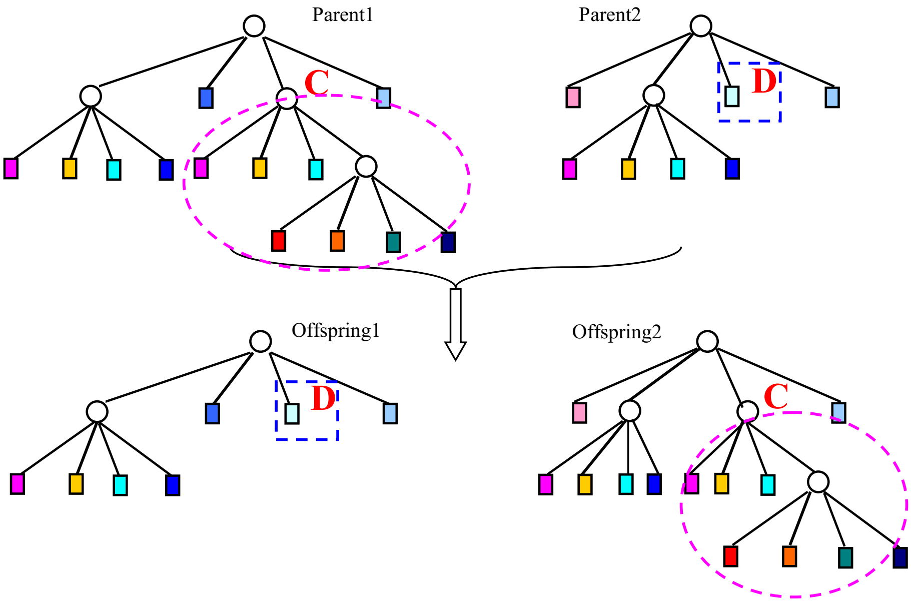


Fig. 3. Crossover swaps two nodes in the same location between two parents. For example, nodes C and D swap and generate two offspring.

Swapping has been widely used in GP [25] to exchange the nodes, and has been shown to be efficient in many applications [26]. In this paper, swapping is only applied to the branches representing the same area in the vegetation map between two parents. Figure 3 shows an example of crossover in the proposed SEA and the generated offsprings. The steps for crossover are described as below:

Step 1: Identify all the nodes in all levels except level 1 with given two parents. Randomly pick specified percentage (swapperc in Table 2) of these nodes from Parent 1 (e.g., C in Figure 3);

Step 2: Determine the node within the same area in Parent 2. (e.g., Node D in Figure 3);

Step 3: Swap all the nodes and leaves that are the descendants of C and D. Go to Step 1 for other parents until all the parents have been randomly chosen based on the crossover probability.

Besides simple swapping, an advanced swapping is proposed in some applications such as the graft crossover [11]. In the graft crossover, different nodes between two parents are identified firstly and the descendants of that node are swapped. This graft crossover can guarantee that two offsprings are different from the parents. In this study, a simple swapping is used, by which some parents with a high fitness value are inherited completely.

There are two benefits for crossover operated on the nodes of a tree in SEA rather than on grids (nodes represent the hierarchical structure of the tree). Firstly, crossover on the nodes can pass on the favorable organizations of vegetation coverage to next generations as discussed in section 3.3. Secondly, it is more computationally efficient for large-scale problems because a crossover on the node can change a relatively big sub-area of a map while a crossover on the grid can only change some pixels of a map.

Unlike the crossover in GP, which can be operated at different branches in the tree or within the same individual, the crossover in SEA swaps nodes representing the same location and is only applied between individuals. This operational restriction in SEA is based on the assumption that a favorable organization in one location may not maintain the fitness at another location.

2.4 Mutation (Splitting and Alternation)

Mutation is important for introducing new information into a population. Conventional mutation is not efficient for large-scale problems because randomly changing some parts of the population without the guidance of spatial patters may be computationally expensive to achieve convergence.

Three criteria are used in designing the mutation operator of the proposed SEA. First of all, the mutation is preferred to be efficiently implemented on the large-scale problem with the help of spatial patterns extracted from the study problem. Second, some randomness must be included to maintain the diversity of the population and balance exploration and exploitation [27-28]. Third, the resultant offspring is required to be a legitimate quadtree to ensure the consistency of the encoding in next generation [11]. Based on these criteria, three operations are used in mutation for

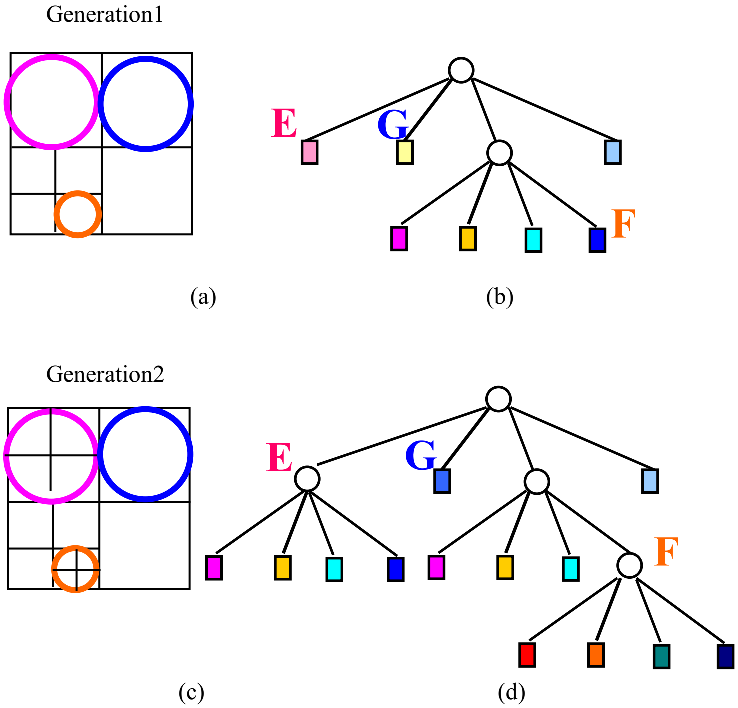


Fig. 4. Spatial mutation has two operations: splitting and alteration. Leaves E and F in (b) have been spitted into four leaves separately as shown in (d). Leaf G has been alternated with another value. And the resolution of the corresponding map has been increased as shown in (c).

image segmentation in the tree-based GA: splitting, merging and alternation [11]. These operations will obviously preserve the quadtree structure. SEA in this paper employs two of those operations: splitting and alternation. Both are operated on the leaves of a tree. Merging is not included in this paper because it is not necessary to decrease the resolution of the spatial map during the optimization. For other spatial optimization problems such as detecting the edges of image segments, merging may be a flexible option of mutation to adjust the resolution.

Similar to the conventional mutation operator, alternation randomly picks some leaves and changes their values. Splitting, which is based on the sensitivity of the leaves to fitness improvement, focuses the computations on interesting subsets of the entire map. Figure 4 shows an example of mutation operations. As discussed in section 3.1, splitting, together with the tree structure, increases computational efficiency and flexibility for large-scale spatial optimization problems.

However, splitting is not totally dependent on sensitivity: some insensitive regions are also selected randomly for splitting to maintain the diversity of the population and reduce the risks of pre-convergence. More diversity of the leaves is also introduced by randomly alternating some leaves based on the specific probability. Both the splitting and alternation can preserve the quadtree structure as Figure 4 shows.

Four parameters control the mutation operation: mutation probability (gam), probability of splitting sensitive leaves (senp), probability of randomly splitting (rsplitp), and a probability for alternating leaves (alterp). Rand is a random number. Figure 5 illustrates the detailed procedures in spatial mutation and Table 2 shows the parameter setting for the testing problem used in this paper.

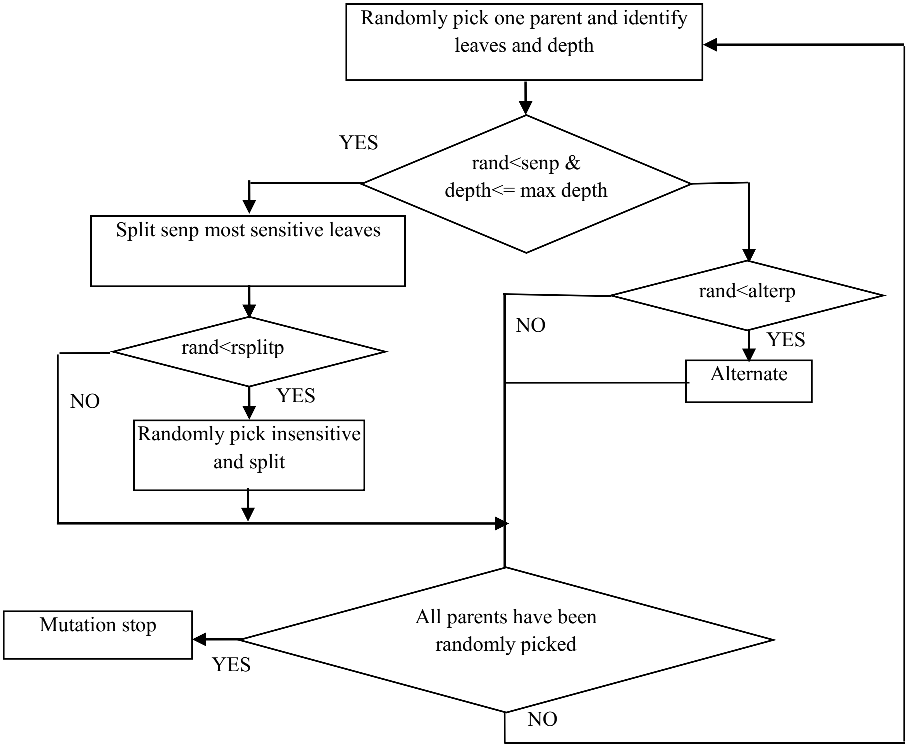

Randomly pick one parent and identify leaves and depth
YES
rand<senp & depth<= max depth
Split senp most sensitive leaves
rand<alterp
NO
YES
Alternate
rand<rsplitp
NO
YES
Randomly pick insensitive and split
Mutation stop
All parents have been randomly picked
YES
NO

Fig. 5. Procedures in the spatial mutation. Note: rand is a random number, senp is the probability of splitting sensitive leaves, rsplitp is a probability of randomly splitting, alterp is a probability for alternating leaves.

3 Illustration Example

As discussed in section 3, SEA is efficient for large-scale spatial optimization problem especially when the computation is beyond the capacity of REA. For small problem when REA can solve, however, SEA is not expected to exceed REA or even takes longer time because the former has extra steps to operate on trees.

To test this hypothesis, a simple 4*4 groundwater model (16 decision variables) is firstly created and 16 grids are assigned with values within the range of [0,0.99]. The “true fitness” is 4.95 for this simple model with enumerations. Then both SEA and REA are employed to solve this simple model and computational time and fitness are compared.

To compare SEA and REA fairly, the groundwater model, the system constraints, and all the shared EA parameters of these two algorithms were set the same for this test. The same conceptual groundwater model was used as the test issue and the groundwater constraints were also the same. The computational experiments are finished with MATLAB Version 7.4 and a Thinkpad laptop of Intel Core 2 Duo CPU and Ram 1.96GB.

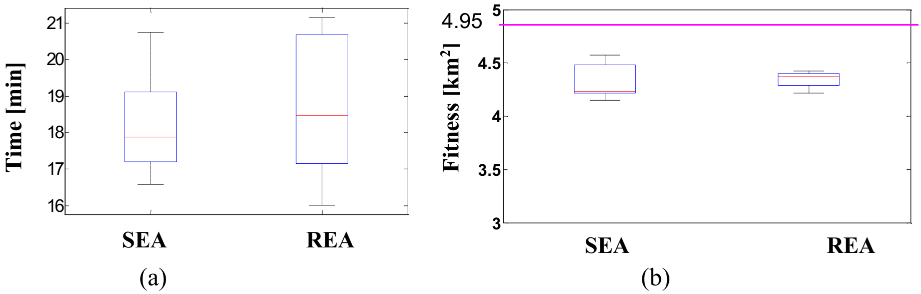


Fig. 6. Comparison of computational time and fitness for small problem with 16 decision variables

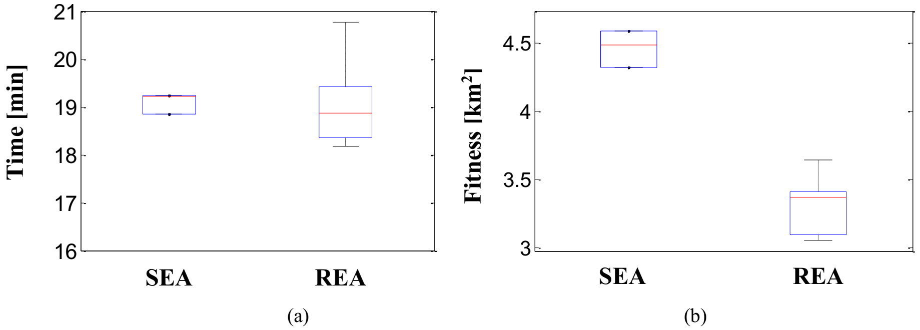


Fig. 7. Comparison of computational time and fitness for a relatively big problem with 64 decision variables

We found that the computational time is very similar between SEA and REA for small problem size with 16 decision variables as Figure 6a shows. In addition, Figure 6b shows SEA has smaller fitness than REA because the former has some approximations when assigning the same values based on patterns. This test result validates the hypothesis that, SEA does not exceed REA for small problem size such as 16 decision variables when REA can solve.

With the same groundwater model, this paper also tested both SEA and REA for a relative bigger problem size with 8*8 grids (64 decision variables). As Figure 7a shows, we found that the compuatational time of SEA and REA is very similar if both algorithms run 100 generations. However, the fitness is quite different: SEA increases the fitness by almost 30% than REA for the test problem with 64 decision variables, which indicates that SEA is more efficient for a bigger problem size. More complete testing and comparison for different problem size is discussed in [23].

4 Discussion and Conclusions

In this paper, the main components of SEA, encoding, initialization and EA operators, have been modified to take advantage of the spatial information so as to solve large-scale spatial optimization problems. The spatial patterns used in crossover, mutation and selection implemented with a tree structure for encoding, distinguishes the SEA from a regular EA. The test example illustrates how SEA encodes the spatial dataset and the SEA procedures.

However, there are some limitations in the application of SEA. First of all, the assumption for the effectiveness of SEA is that there exists spatial patterns in the spatial dataset of decision variables. If the pattern does not exist, the decision map will be essentially based on the manipulation of grids. SEA does not provide an accurate solution. In particular, when the spatial dataset has a checkerboard pattern [24] and does not have any neighboring pattern, the data volume of SEA cannot recognize it. Secondly, the accuracy of SEA solution depends on the resolution of the map (i.e., how much detail the map shows) For a small scale problem like the illustration example with only 16 decision variables, REA is feasible and more accurate as shown in Figure 6.The developed SEA is motivated by the tree-based EA developed by Gong and Yang [11] and Laszlo and Mukherjee [12]. But there are two major differences between the developed SEA and their methods: (1) They didn't incorporate spatial patterns in the selection operator; (2) they used an energy function as an optimization objective while the developed SEA used a management objective to solve the problems of resources allocation. Many scientists used tree-based GA and spatial dataset in the field of image processing while few scientists adopted this idea to solve a spatial optimization problem in the field of water resources planning and management. The challenge is to customize EA operators and constraints to better accommodate the characteristics of a specific problem.

References

1. Loonen, W., Heuberger, P., Kuijpers-Linde, M.: Spatial optimization in land-use allocation (2007)
2. Grimm, V., Revilla, E., Berger, U., Jeltsch, F., Mooij, W.M., Railsback, S.F., Thulke, H., Weiner, J., Wiegand, T., DeAngelis, D.L.: Pattern-Oriented Modeling of Agent-Based Complex Systems: Lessons from Ecology. Science 310, 987–991 (2005), doi:10.1126/science.1116681.
3. McKinney, D.C., Lin, M.D.: Genetic algorithm solution of groundwater management models. Water Resource Research 30(6), 1897–1906 (1994)
4. Hilton, A.B.C., Culver, T.B.: Constraint handling for genetic algorithms in optimal remediation design. Journal of Water Resources Planning and Management 126(3), 128–137 (2000)
5. Schütze, N., Paly, M., Schmitz, G.: Optimal open-loop and closed-loop scheduling of deficit irrigation systems. Journal of Hydroinformatics 14(1), 136–151 (2012)
6. Krzanowski, R.M., Raper, J.: Spatial Evolutionary Modeling. Oxford University Press, Oxford (2001)
7. Openshaw, S.: Developing automated and smart spatial pattern exploration tools for geographical information systems applications. The Statistician 44(1), 3–16 (1995)

8. Openshaw, S.: Neural network, genetic, and fuzzy logic models of spatial interaction. Environment and Planning A 30, 1857–1872 (1998)
9. Xiao, N.: A unified conceptual framework for geographical optimization using evolutionary algorithms. Annals of the Association of American Geographers 98(4), 795–817 (2008)
10. Brooks, C.: A genetic algorithm for designing optimal patch configurations in GIS. International Journal of Geographical Information Science 15(6), 539–559 (2001)
11. Gong, M., Yang, Y.: Quadtree-based genetic algorithm and its applications to computer vision. Pattern Recognition 37, 1723–1733 (2004)
12. Laszlo, M., Mukherjee, S.: A genetic algorithm using hyper-quadtrees for low-dimensional k-means clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(4), 533–543 (2006)
13. Wang, S., Armstrong, M.P.: A quadtree approach to domain decomposition for spatial interpolation in Grid computing environments. Parallel Computing 29, 1481–1504 (2003)
14. Xiao, N., Bennett, D.A., Armstrong, M.P.: Using evolutionary algorithms to generate alternatives for multiobjective site search problems. Environment and Planning A 34(4), 639–656 (2002)
15. Cao, K., Batty, M., Huang, B., Liu, Y., Yu, L., Chen, J.: Spatial multi-objective land use optimization: extensions to the non-dominated sorting genetic algorithm-II. International Journal of Geographical Information Science 25(12), 1949–1969 (2011)
16. Fotakis, D., Sidiropoulos, E.: A new multi-objective self-organizing optimization algorithm (MOSOA) for spatial optimization problems. Applied Mathematics and Computation 218, 5168–5180 (2012)
17. Laszlo, M., Mukherjee, S.: A genetic algorithm using hyper-quadtrees for low-dimensional k-means clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(4), 533–543 (2006)
18. Holland, J.H.: Adaptations in Natural and Artificial Systems. University of Michigan Press, Ann Arbor (1975)
19. De Jong, K.A.: Evolutionary computation: where we are and where we're headed. Fundamenta Informaticae 35, 247–259 (1998)
20. Fogel, D.B.: Evolutionary Computation: Toward a New Philosophy of Machine Intelligence. IEEE Press, Piscataway (2000)
21. Cova, T.J., Church, R.L.: Contiguity constraints for single-region site search problems. Geographical Analysi. 32(4), 306–329 (2000)
22. Wang, D., Cai, X.: Irrigation Scheduling-Role of Weather Forecasting and Farmers. Journal of Water Resources Planning and Management 135(5), 364–372 (2009)
23. Wang, J., Cai, X., Valocchi, A.J.: Spatial evolutionary algorithm (SEA) for optimizing a large-scale irrigation pumping strategy. In: INFORMS Annual Meeting, Charlotte, NC (2011)
24. Samet, H.: The Design and Analysis of Spatial Data Structures. Addison-Wesley, New York (1990)
25. Koza, J.R.: Genetic Programming: On the Programming of Computers by Means of Natural Selection. MIT Press, Cambridge (1992)
26. Yang, Y.C.E., Cai, X., Herricks, E.E.: Identification of hydrologic indicators related to fish diversity and abundance: A data mining approach for fish community analysis. Water Resources Research 44, W04412 (2008), doi:10.1029/2006WR005764.
27. Holland, J.H.: Genetic algorithms. Scientific American 267(1), 66–72 (1992)
28. Sefrioui, M., Périaux, J.: A hierarchical genetic algorithm using multiple models for optimization. In: Deb, K., Rudolph, G., Lutton, E., Merelo, J.J., Schoenauer, M., Schwefel, H.-P., Yao, X. (eds.) PPSN 2000. LNCS, vol. 1917, pp. 879–888. Springer, Heidelberg (2000)

Part III
Wearable Computing and Intelligent Data Hiding

Block-Based Colour Image Steganography Using Smart Pixel-Adjustment

Ching-Yu Yang[1] and Wen-Fong Wang[2]

[1] Dept. of Computer Science and Information Engineering,
National Penghu University of Science and Technology, Penghu, Taiwan
chingyu@npu.edu.tw
[2] Dept. of Computer Science and Information Engineering,
National Yunlin University of Science and Technology, Yunlin, Taiwan
wwf@yuntech.edu.tw

Abstract. By adjusting the pixel-value of a host block, we design an effective steganographic method for color images. Specifically, based on a smart pixel-adjustment policy with two averages of the block, a secret message can be embedded in a host image without arising visual distortion. Experiments indicate that the perceived quality generated by the proposed method is good while the payload is larger than existing techniques. Moreover, the proposed method has a merit of maintaining a certain degree of robustness. Namely, the marked images generated by our method are tolerant of manipulations such as color quantization, equalized, edge sharpening, inversion, JPEG, JPEG2000, noise additions, pixel-truncation, winding, and zigzagging. This robustness is rarely found in the traditional techniques for color image steganography.

Keywords: Data hiding, color image steganography, smart pixel-adjustment policy.

1 Introduction

Thanks to the economic, fast broadband services, and a variety of interesting or useful applications developed on the mobile devices, people is capable of easily sharing (and exchanging) their resources on the internet. However, data can be eavesdropped, falsified, and tampered with during transmission. Data hiding techniques are gradually attracted people and being used to against the above challenge. Several applications of data hiding can be found in protecting intellectual property right, content authentication, and copyright protection [1,2]. In general, data hiding can be divided into two categories: steganography and digital watermarking. The key feature of watermarking approaches [3,4] is robustness performance. Namely, most of the watermarked images are tolerant of manipulations, such as compression, cropping, noise-addition, and quantization attacks. However, most of the conventional watermarked methods introduce a limited size of payload. One of the main goal of the steganographic methods [5,6] is the providing of high storage for covert

H. Sun et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 329, DOI: 10.1007/978-3-319-12286-1_15

communications between (or among) the parties. How to maintain a good perceived quality while providing a high-capacity are an important issue for the design of an image steganography. Since color images are more attractive than gray-level ones and are commonly circulated around the world. Several researchers have presented data hiding techniques for color images [4,5,6,7,8]. Yang [5] used the radius weighted mean (RWM) and presented a steganographic method for color images. Based o the RWM-decision policy with the least significant bit (LSB) substitution technique, data bits are effectively embedded in a host color images. Simulations showed that the resultant payload is larger than existing techniques while the perceived quality is not bad. Based on edge detection and an efficient hiding technique, Ioannidou et al. [6] proposed a novel image steganography. Experimental results indicated that their peak-to-signal ratio (PSNR) performance is good with a limited size of payload. Moreover, the overhead information of this technique is significant. Namely, two auxiliary files have to be sent to the receiver to extract secret bit. Based on pixel value differencing (PVD) scheme, Mandal and Das [7] suggested a color image steganography in spatial domain. To further promote security and avoid overflow issue, the proposed scheme embedded different number of data bits in different pixel component. Experiments demonstrated that the resultant payload size is large while the perceived quality is good.

In this paper, we propose a steganographic method for color images based on a smart pixel-adjustment policy with two averages of the block. The rest of the paper is organized as follows. The proposed bit-embedding and bit-extraction techniques are described in Section 2. Experimental results are demonstrated in Section 3. Finally, a conclusion is summarized in Section 4.

2 The Proposed Method

First, the R-/G-/B-plane of a host color image are separately extracted. Then, each plane is divided into a series of non-overlapping n×n blocks. A secret message is subsequently embedded in the host blocks of each plane, respectively. Since the proposed method uses the same idea to embed data bits in the blocks of R-/G-/B-plane of a host color image. For clarity, only data hiding of the R-plane is presented here, while that of G-/B-plane are omitted. The details of the procedure for the proposed bit-embedding and bit-extraction in the R-plane are described in the following sections.

$(r_{i-1,,j-1})$	$(r_{i-1,j})$
$(r_{i,j-1})$	$(r_{i,,j})$

Fig. 1. A 2×2 host block of the R-plane

2.1 Bit-Embedding

Without loss of generality, let $H_k = \{r_{i-1,j-1}, r_{i-1,j}, r_{i,j-1}, r_{i,j}\}$ be the *k*th host block divided from R-plane of a RGB color space, as shown in Fig. 1 (when *n*=2), and $H_k = \hat{H} \cup \tilde{H}$ with $\hat{H} = \{r_{i-1,j-1}, r_{i,j}\}$ and $\tilde{H} = \{r_{i-1,j}, r_{i,j-1}\}$ Also let $M_{\hat{H}} = \lfloor (r_{i-1,j-1} + r_{i,j})/2 \rfloor$ and $M_{\tilde{H}} = \lfloor (r_{i-1,j} + r_{i,j-1})/2 \rfloor$ be the two average values of the block. The proposed bit-embedding is described in the following algorithm.

Algorithm 1. Hiding data bits in the R-plane of a color image.

```
Input: A R-plane image $S_R$, a secret message $T$, and two control parameter $\tau_r$ and $\lambda$.
Output: A marked image $S'_R$.
Method:
Step 1. Input a block $H_k$, which derived from $S_R$. If the end of input is
    encountered, then proceed to Step 6.
Step 2. Compute the offset $\varphi$ of the block, i.e., $\varphi = M_{\hat{H}} - M_{\tilde{H}}$.
Step 3. Obtain one data bit $\delta$ from $T$, and perform the following sub-steps:
Step 4. If $\delta$ =1, then do the following sub-steps:
    Step 4.1. If both conditions of $\varphi < 0$ and $\varphi \geq -\tau_r$ are satisfied, then do nothing,
        which means the block "carries" data bit 1 without altering the pixels' value,
        and proceed to Step 1.
    Step 4.2. If $\varphi < -\tau_r$, we repeatedly increase the pixel-value in $\hat{H}$ and decrease
        pixel-value in $\tilde{H}$ by the $\lambda$ value each time, simultaneously, until either the
        conditions of $\varphi < 0$ and $\varphi \geq -\tau_r$ are satisfied or times $\tau_r$ is encountered.
    Step 4.3. If both conditions of $\varphi < 0$ and $\varphi \geq -\tau_r$ are satisfied, then proceed to
        Step 1; otherwise, undo the above pixel-value adjustment, mark the block as
        a 'skipped block', and proceed to Step 1.
    Step 4.4. If $\varphi > \tau_r$, we repeatedly increase the pixel-value in $\tilde{H}$ and decrease
        pixel-value in $\hat{H}$ by the $\lambda$ value each time, simultaneously, until either the
        conditions of $\varphi < 0$ and $\varphi \geq -\tau_r$ are satisfied or times $\tau_r$ is encountered.
    Step 4.5. If both conditions of $\varphi < 0$ and $\varphi \geq -\tau_r$ are satisfied, then proceed to
        Step 1; otherwise, undo the above pixel-value adjustment, mark the block as
        a 'skipped block', and proceed to Step 1.
```

Step 5. If δ =0, then do the following sub-steps:

Step 5.1. If both conditions of $\varphi \leq \tau_r$ and $\varphi \geq 0$ are satisfied, then do nothing, which means the block "carries" data bit 0 without altering the pixels' value, and proceed to Step 1.

Step 5.2. If $\varphi > \tau_r$, we repeatedly increase the pixel-value in $\tilde{H}$ and decrease pixel-value in $\hat{H}$ by the λ value each time, simultaneously, until either the conditions of $\varphi \leq \tau_r$ and $\varphi \geq 0$ are satisfied or times τ_r is encountered.

Step 5.3. If both conditions of $\varphi \leq \tau_r$ and $\varphi \geq 0$ are satisfied, then proceed to Step 1; otherwise, undo the above pixel-value adjustment, mark the block as a 'skipped block', and proceed to Step 1.

Step 5.4. If $\varphi < 0$, we repeatedly increase the pixel-value in $\hat{H}$ and decrease pixel-value in $\tilde{H}$ by the λ value each time, simultaneously, until either the conditions of $\varphi \leq \tau_r$ and $\varphi \geq 0$ are satisfied or times τ_r is encountered.

Step 5.5. If both conditions of $\varphi \leq \tau_r$ and $\varphi \geq 0$ are satisfied, then proceed to Step 1; otherwise, undo the above pixel-value adjustment, mark the block as a 'skipped block', and proceed to Step 1.

Step 6. Stop.

To avoid overflow during the pixel adjustment, the increment operation is bypassed to the pixels of which value being greater than (255-τ_r). Similarly, to avoid underflow during the reducing procedure, the decrement is bypassed to the pixels of which value being less than or equal to τ_r. Notice as well the number of skipped blocks is zero when a host block of size 2×2 (or 3×3) was used in our method. Namely, no overhead such as bitmap is required by the proposed method.

2.2 Bit-Extraction

Let $H'_k = \{r'_{i-1,j-1}, r'_{i-1,j}, r'_{i,j-1}, r'_{i,j}\}$ be the *k*th hidden block divided from R-plane of a marked image, $H'_k = \hat{H}' \cup \tilde{H}'$ with $\hat{H}' = \{r'_{i-1,j-1}, r'_{i,j}\}$ and $\tilde{H}' = \{r'_{i-1,j}, r'_{i,j-1}\}$ Also let $M_{\hat{H}'}$ and $M_{\tilde{H}'}$ be the two average values of the block. The procedure of bit-extraction is much simpler than that of bit-embedding. The algorithm of the proposed bit-extraction is specified in the following steps.

```
Input: A marked image $S'_R$, and the control parameter $\tau_r$.
Output: A secret message $T$.
Method:
Step 1. Input a hidden block $H'_k$ , which derived from $S'_R$. If the end of input is
    encountered, then proceed to Step 4.
Step 2. Compute the offset of the block $\varphi = M_{\hat{H}'} - M_{\tilde{H}'}$.
Step 3. If both conditions of $\varphi < 0$ and $\varphi \geq -\tau_r$ are satisfied, then data bit 1 is
        extracted, otherwise, data bit 0 is extracted, and go from Step 1.
Step 4. Assemble the extracted data bits to form a secret message $T$.
Step 5. Stop.
```

As mentioned previously, a secret message is separately embedded in the three planes of a color host image. Namely, the bit-embedding sequence follows the order of R-plane, G-plane, and R-plane,

3 Experimental Results

Several 512×512 color images were used as host images. Each RGB pixel of the host images is represented by 24 bits, 8 bits per component. The size of the test binary watermark is 444×444. The block size is 2×2. The marked images generated by the proposed method are depicted in Fig. 2. Notice that each host image equipped with a variety of control parameters ($\tau_r/\tau_g/\tau_b$) in the R/G/B-plane. The integer λ is set to 1. For example, a set of control parameters, τ_r=19, τ_g=32, and τ_b=27 was used to introduce the marked image *Lena*. No skipped blocks were generated. Namely, the optimal number of hidden bit for each marked images is $(512\times512)/2\times2\times3 = 196{,}608$ bits. From the figure we can see that the perceived quality is good. No false colors appeared in the figures. Their corresponding PSNR is given in Table 1. In addition, the performance of the proposed method using a host block of size 3×3 is included. From Table 1 we can see that the payload for a block of size 2×2 is larger than that of the blocks with size of 3×3, while the average PSNR of the former is slightly smaller than that of the latter. The PSNR is defined by

$$PSNR = 10\times\log_{10}\frac{255^2}{MSE} \tag{1}$$

with $MSE = (\sum_{i=1}^{MN}\left[(r_i-\hat{r}_i)^2 + (g_i-\hat{g}_i)^2 + (b_i-\hat{b}_i)^2\right])/3MN$. Here (r_i, g_i, b_i) and $(\hat{r}_i, \hat{g}_i, \hat{b}_i)$ denote the RGB pixel values of the host image and the marked image.

(a) (b) (c) (d) (e) (f)

Fig. 2. The marked images generated by the proposed method using various control parameters (τ_r/ τ_g/ τ_b) in host images. (a) Lena (19/32/27), (b) Jet (20/37/23), (c) Peppers (28/39/33), (d) Tiffany (40/77/69), (e) Splash (14/66/34), (f) House (5/6/7), (g) Couple (9/10/9) and (h) Car-House (86/40/24).

(g) (h)

Fig. 2. (*continued*)

Table 1. PSNR and payload comparison between the proposed method using the blocks with different size

Images	Block size	
	2×2	3×3
Lena	41.58/196,608	43.65/86,700
Jet	43.73/196,608	43.10/86,700
Peppers	39.43/196,608	40.20/86,700
Tiffany	40.57/196,608	42.84/86,700
Splash	44.86/196,608	45.36/86,700
House	49.17/196,608	47.10/86,700
Couple	49.54/196,608	47.96/86,700
Car-House	41.34/196,608	42.04/86,700
Average	43.78/196,608	44.03/86,700

Table 2. PSNR (dB) and payload (bit) comparison between various methods

Images	Methods			
	Yang's scheme [5]	Ioannidou et al.'s approach [6]+	Mandal and Das's technique [7]	Our Method*
Lena	45.32/171,200	45.12/30,987	42.26/145,787	41.58/196,608
Baboon	46.44/160,300	-	38.44/144,916	33.29/196,608
Jet	44.90/199,500	-	42.60/145,648	43.73/196,608
Peppers	46.20/181,300	44.45/~31kb	42.28/145,995	39.43/196,608
Car-House	47.61/110,870	-	41.41/145,374	41.34/196,608
Splash	45.24/145,900	-	42.86/146,732	44.86/196,608
Sailboat	46.61/166,780	-	40.66/143,278	47.41/196,608
Tiffany	42.12/165,720	43.87/~31kb	-	40.57/196,608
House	45.48/175,440	45.12/~31kb	-	49.17/196,608
Average	45.55/164,112	44.64/~31kb	41.50/145,390	43.80/196,608

+ The method uses the Laplacian or fuzzy edge detector without using random number generator.

* The block size is 2×2.

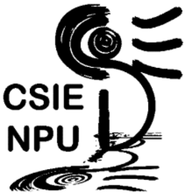
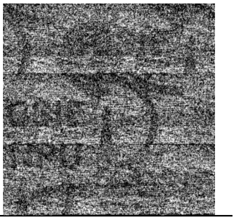
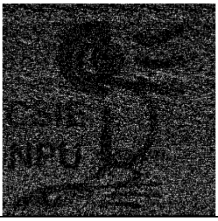
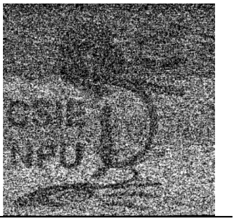
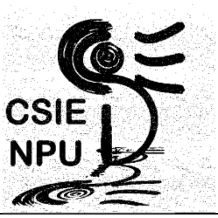
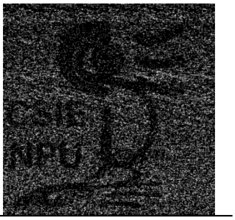
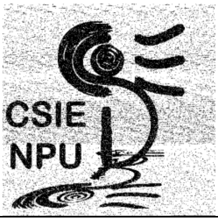
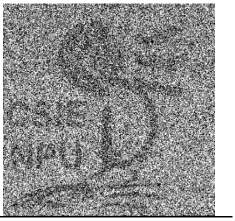
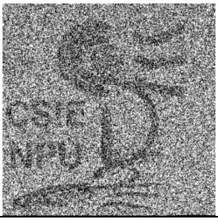
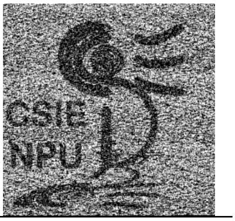
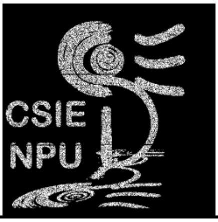
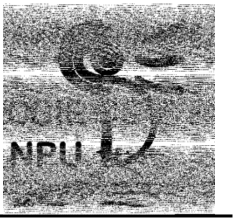
Performance comparison between the proposed method and existing schemes: Yang's scheme [5], Ioannidou et al.'s approach [6], and Mandal and Das's technique [7] is given in Table 2. It can be seen that the proposed method provides the largest payload among these compared methods while the average PSNR of our method is still larger than that of Mandal and Das's technique [7]. The average PSNR for the proposed method is slightly less than that of Yang's scheme [5] and Ioannidou et al.'s approach [6]. However, the hiding capacity provided by our method is about 6 times larger than that provided by Ioannidou et al.'s approach [6].

To demonstrate the proposed steganographic method shares a certain degree of robustness performance, the marked images are tested by a variety of attacks. The extracted watermarks and their bit correct ratio (BCR) are given in Table 3. The size of an input watermark is 443×443.The tested marked image is generated by the proposed method using τ_r=19, τ_g=32, τ_b=27, and λ=1, respectively, on image Lena. The BCR is defined by

$$BCR = \left(\sum_{i=0}^{ab-1} \overline{w_i \oplus \tilde{w}_i} \Big/ a \times b \right) \times 100\%, \tag{2}$$

where w_i and $\tilde{w}_i$ represent the values of the original watermark and the extracted watermark respectively, as well as the size of a watermark is $a \times b$. The BCR for an

Table 3. The survived watermarks extracted from the marked images which undergone various manipulations

Attacks	Survived watermarks	Attacks	Survived watermarks
Null Attack BCR = 100%		JPEG2000 (CR=9.97) BCR=50.83%	
Color quantization (8-color) BCR=35.68%		JPEG (QF=90) BCR=51.71%	
Edge Sharpening BCR=97.57%		Truncation* BCR=35.88%	
Equalized BCR=90.99%		Uniform noise (6%) BCR =60.45%	
Gaussian noise (3%) BCR =59.01%		Winding BCR =60.69%	
Inversion BCR=7.51%		Zigzagging BCR =66.38%	

* The last five-bit of the stego-pixels were purposely truncated.

extracted watermark is 100% if a marked image is not manipulated. Table 3 shows that most extracted watermarks are recognized. Note that the BCR of the survived watermark extracted from a marked image, which had undergone inversion attack, is only 7.51% and is still recognizable. In addition, the extracted watermarks are recognized when the marked images were compressed by JPEG2000/JPEG with a compression ratio (CR) about 10 and 4, respectively. Similar performance can be found in the marked images which manipulated by colour quantization, pixel-truncation, winding and zigzagging. Moreover, the survived watermarks are identified after the marked images undergone Uniform/-Gaussian-noise addition attacks.

4 Conclusion

In this paper, a simple but effective method for color image steganography was proposed. By using a smart pixel-adjustment technique in accordance with the two average values of the block, data bits can be effectively embedded in a host image with no color distortion. Our simulations confirm that the visual quality introduced by the proposed method is good while the payload for the proposed method is better than existing techniques. Additionally, the proposed method shares a certain degree of robustness. The marked images generated by our method are tolerant of manipulations such as color quantization, equalized, edge sharpening, inversion, JPEG, JPEG2000, noise additions, pixel-truncation, winding, and zigzagging. Most of the conventional steganographic approaches rarely possess this robustness performance.

References

1. Cox, I.J., Miller, M.L., Bloom, J.A., Fridrich, J., Kalker, T.: Digital Watermarking and Steganography, 2nd edn. Morgan Kaufmann, MA (2008)
2. Phadikar, A.: Data Hiding Techniques and Applications Specific Designs. LAP LAMBERT Academic Publishing, Saarbrucken (2012)
3. Noriega, R.M., Nakano, M., Kurkoski, B., Yamaguchi, K.: High payload audio watermarking: toward channel characterization of MP3 compression. Journal of Info. Hiding and Multimedia Signal Processing 2(2), 91–107 (2011)
4. Niu, P.P., Wang, X.Y., Yang, Y., Lu, M.Y.: A novel color image watermarking scheme in nonsampled contourlet-domain. Expert System with Applications 38, 2081–2098 (2011)
5. Yang, C.Y.: Use of radius weighted mean to hide data in colour images. In: The 5th IET Int. Conf. on Ubi-Media Computing, Xining, China, August 16-18 (2012)
6. Loannidou, A., Halkidis, S.T., Stephanides, G.: A novel technique for image steganography based on a high payload method and edge detection. Expert System with Applications 39, 11517–11524 (2012)
7. Mandal, J.K., Das, D.: Colour image steganography based on pixel value differencing in spatial domain. Int. Journal of Information Science and Techniques 2, 83–93 (2012)
8. Eielinska, E., Mazurczyk, W., Szczypiorski, K.: Trends in steganography. Communications of the ACM 57, 86–95 (2014)

A Sport Recognition Method with Utilizing Less Motion Sensors

Wen-Fong Wang[1], Ching-Yu Yang[2], and Ji-Ting Guo[1]

[1] National Yunlin University of Science & Technology,
Department of Computer Science & Information Engineering,
Douliu, Yunlin, 640 Taiwan
{wwf,m10017006}@yuntech.edu.tw

[2] Department of Computer Science and Information Engineering, National Penghu
University of Science & Technology, Makung, Penghu, 880 Taiwan
chingyu@npu.edu.tw

Abstract. In this study, we propose a recognition method in ball games using no more than two triaxial accelerometers on the user's front arm and upper arm to track motion data. To produce effective features for classifying ball games' postures, the motion data is processed by our method, which includes a median filter, a duplication removal algorithm, and an algorithm of feature extraction. Subsequently, the produced features are recognized by a support vector machine scheme for sports with single-handed swings like tennis, badminton, and ping pong. The research result in this investigation can help the athlete training of the above mentioned sports. Experimental results showed that the precision rate of the proposed method for recognizing postures in a single-handed swing achieves 95.67%.

Keywords: Accelerometry, single-handed swing, swing features, feature extraction, swinging posture recognition.

1 Introduction

In previous study, the detection of human activity in sports or daily lives is achieved by setting up sensory devices in the living environment or worn on the body of users. To detect human activity based on video or image sensors, it has weaknesses such as the processing complexity of images or video. However, the most unacceptable weakness is the violation of personal privacy. In addition, most of the researches feature numerous sensory devices, resulting in inconvenience to the users, high amount of system complexity, and cost issues. Thus rendering devices mentioned above uneasy to be applied in daily lives. The application of triaxial accelerometers circumvents the above issues since it is easy for applications, low cost, and low power consumption.

Recently, researchers used triaxial accelerometers to recognize the basic postures and movements of a human body, including walking, running, standing, and crouching, with a precision rate of above 90% [1] [2] [3] [4]. However, the joints of a human body constitute at least 244 dimensions of freedom for bodily

H. Sun et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 329, DOI: 10.1007/978-3-319-12286-1_16

movements [5], and a single triaxial accelerometer is thus unable to produce data precise enough to determine the complex movements of the human body.

Certain methods, which combine environmental information with sensors worn on a user, utilize the sensors in the surrounding environment to detect the position of the user and then determine the user's actions using data from the user's triaxial accelerometers [6] [7]. With these methods, however, the user's movement is limited to a pre-determined space. Furthermore, the installation of such facilities will result in high system complexity and installation costs, and thus make this method practically unrealistic. In more recent research, the proposed schemes in [8] and [9] consisted of placing multiple triaxial accelerometers on a subject, using the triaxial accelerometers as a primary sensing device in combination with secondary devices such as Force Sensing Resistors, to measure muscle contractions. These researches indicated that in order to perceive complex postures, numerous sensors must be assigned to relevant areas to collect information to allow the algorithms to correctly determine the type of a posture the user is taking.

To identify the data collected from the motion sensors, classification algorithms must be taken into account. In previous research, classification algorithms, including decision trees, finite state machines (FSMs), activity track algorithms, and support vector machines (SVMs), have been mentioned. The studies on the decision trees [10] [11] were based on the users' basic postures, such as standing, walking, and bending. The prediction accuracy of such studies can be up to 90%. However, there is one drawback while applying this method to posture recognition. If the number of postures to be recognized were increased, the reconstruction of decision trees would be processed again and cause a lot of efforts. In FSMs, a state can be used to represent a posture in a continuous activity for recognizing a sequence of posture changes. In [12], the study indicated that one can precisely identify a user's current posture using a FSM algorithm. This method can accurately recognize postures with clear distinctions; for example, the postures involved in a standing up movement includes sitting and standing postures. However, as it comes to complicated postures without clear distinctions, the accuracy will greatly decrease.

An activity track algorithm encodes and breaks down a series of postures and locates the most suitable template module from the series. A tracked template module represents a classified action that is recognized by the template [13]. The advantage of this method is that with sufficient motion sensors to collect users' motion data for producing template modules, the constructed modules would be suitable for most of user population. Furthermore, the problem with users' corelation does not exist. However, the flaw in this method is similar to that in the FSM algorithms; neither method can clearly identify the distinctions between postures. To acquire enough data for the encoding, lot of sensors have to be mounted on the users, resulting in system complexity and an increase in cost.

For SVMs, it is a type of neural networks that are sorted in supervised learning. Neural networks can achieve excellent accuracy with proper training.

Previous research studies that used this method achieved an accuracy rate of 95% or more, which makes SVM an accurate and practical method [14] [15].

Previous posture capturing studies that used the triaxial accelerometers are confined to basic postures, despite the fact that users' daily postures are not limited only to walking, running, standing, or sitting. Similar equipment, such as pedometers, can classify the user's walking data. Advanced devices embedded with a global positioning system (GPS) can even record the distance and route of a user [16]. In the studies of using multiple motion sensors, the users are mostly restricted in a certain space to perform their postures; the findings of such an approach, however, cannot be broadly applied in the context of people's daily lives. Nonetheless, the possibility of using triaxial accelerometers to study swing posture's recognition still has a great opportunity for research and applications. Thus, we try to propose a posture recognition method, which is not restricted by testing environment, but by using only one or two triaxial accelerometers.

In this study, we propose a sport recognition method to detect the postures such as tennis, badminton, and ping pong. To track motion data from the arm of a subject, two triaxial accelerometers were installed on the subject's front arm and upper arm. Then, the motion data is used by a posture recognition system developed to analyze the features in a single-handed swing (for example, tennis, badminton, and ping pong). The remainder of this paper is organized as follows. The system flowchart and the definition of features are outlined in Section 2. The specifications of the proposed methods, which include data collection, feature extraction, and criteria for performance, are given in Section 3. Section 4 specifies the design of experiments and experimental results. Finally, Section 5 gives the conclusions.

2 System Architecture

2.1 System Flowchart

Badminton, tennis, and ping pong are the three most common sports that feature a single-handed swing. The swing is composed by continuously moving one's upper arm and front arm to achieve maximum speed acceleration upon hitting the ball. Therefore, if we track and record the acceleration variations of these two parts of the arm, we can identify the posture of a subject hitting the ball. In our system, the measuring device is attached around the outer side of the subject's arm. This measuring device continuously collects the output of a triaxial accelerometer.

The flowchart for processing the acceleration data of each dimension is shown in Figure 1. First, the X-axis acceleration data is extracted and passed through a median filter (with a window size of 5) so as to exclude the noise data from the source. The feature extraction is then to acquire the critical information needed to identify the posture. In addition, the appearing time of each peak value, which is a useful hint when determining the position of arms, is recorded. On the other hand, the Y-axis acceleration data is extracted for verifying the peak appearance time of X-axis acceleration data in order to extract the Y-axis

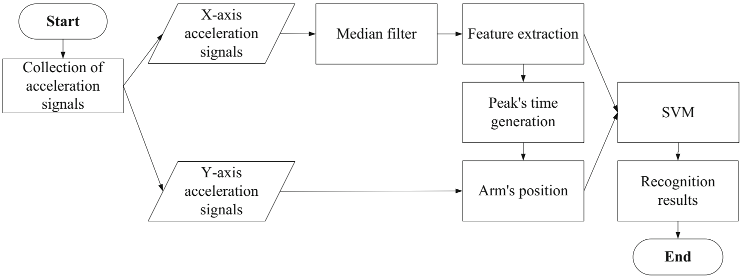


Fig. 1. System flowchart

acceleration data of each time segment for determining the arm's position. After feature extraction and arm's position identification, all feature data is sent to a trained SVM to acquire the features and identify the subject's posture.

2.2 Feature Composition

Once the postures of a swing have been extracted, we can clearly identify the segmented postures, including the readying posture with arms stretched to the back, the swinging posture when hitting the ball, and the follow-through movements after the ball is hit. The ball hitting posture has the highest instant acceleration values of the three postures. These three postures are generated sequentially in a short period of time. Thus, if the three postures and their order were extracted strictly, we can effectively increase the precision rate of identifying each posture. To aid the postures in recognition, the features picked from the three posture segments are as follows: BHP, HP, and AHP (as shown in Table 1). However, with the three mentioned features, it is not possible to identify a subject's instant posture. Therefore, additional five features must be added in between the three main features to support recognition. These features are as follows: DAB, DHA, DHB, SHB, and SHA. Detailed definitions of the features are given in Table 1. Figure 2 shows an ideal collection of features.

Table 1. Feature definition

Feature Name	Definition
HP	The highest peak value of a feature set
BHP	The neighboring peak value before HP
AHP	The neighboring peak value after HP
DAB	The distance between BHP and AHP
DHA	The distance between HP and AHP
DHB	The distance between HP and BHP
SHB	The difference between HP and BHP
SHA	The difference between HP and AHP

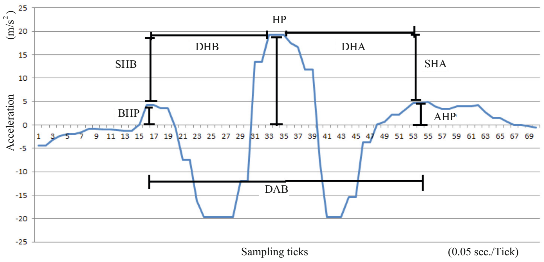


Fig. 2. Ideal collection of features for a posture recognition

3 Proposed Methods

3.1 Sensor Device and Data Collection

In this work, our sampling rate is set to 50Hz; that is, each second is divided into fifty sampling ticks. The devices collect acceleration signal's data from the following three dimensions: X, Y, and Z dimensions. The measuring devices are attached on the upper arm and front arm are shown in Figure 3. As the arm lifts horizontally, the direction of the Y-axis becomes parallel to the subject's neck and points to the top of the head. The X-axis becomes parallel to the elbow and points to the shoulder. The Z-axis becomes parallel to the subject's outer side of the body. In the initial condition, since the Y-axis is perpendicular to the ground surface, the acceleration data collected is equal to $9.8m/s^2$ of gravity. The other two axes do not display significant changes, and the acceleration data for both of the axes are almost zero. The recording device can individually record the time of each entry for reference in statistical analysis. During the experiments, subjects are asked to perform one of the following sports: badminton, tennis, or ping pong. The timer is set to 5 minutes for each subject, and the subject is asked to perform swings in a normal condition.

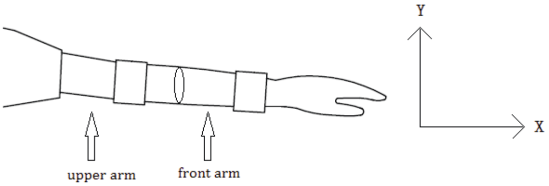


Fig. 3. Illustration of attaching the measuring devices

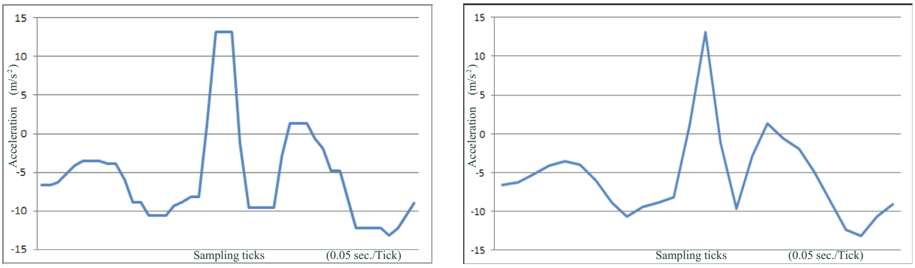
3.2 Feature Extraction

According to the attachment of the sensing devices, the acceleration along the X-axis in the swinging posture is affected the most significantly. Therefore, the eight features listed in Table 1 along the X-axis are extracted as an important reference when identifying the posture. Yet the original data still displays noise data.

To eliminate the noise data, we utilize a median filter to achieve the goal. More specifically, let $\boldsymbol{X} = \{x_0, x_1, \cdots, x_{(N-1)}\}$ be an one-dimensional array of original signals of length N, $\boldsymbol{Y} = \{y_0, y_1, \cdots, y_{(N-1)}\}$ an one-dimensional signal array after filtering, and $S_k = (s_{(i-\frac{M-1}{2})}, \cdots, s_k, \cdots, s_{(i+\frac{M-1}{2})})$ a signal vector of length M (where M is an odd integer) for finding a signals' median in the vector. The median filter is defined as follows:

$$y_i = \mathbf{median}_i(S_k), \tag{1}$$

where $i = 0 \ldots (N\text{-}1)$, $M < N$, and $(i - \frac{M-1}{2}) \leq k \leq (i + \frac{M-1}{2})$. The size of the signal vector (i.e. M) can affect the effectiveness of noise elimination. As M is set to 5, it can eliminate the possible noise values without affecting the completeness of the data. Actually, setting M to 5 is a preferred value since setting the value over 5 ruins the required data and setting it below 5 cannot eliminate the noise.



(a)

(b)

Fig. 4. The motion signals – (a) with duplicated signal values and (b) after eliminating duplicated medians

From Figure 4(a), we can see that the noise issue has been significantly improved. However, numerous duplicated signal values are created due to applying the median filter. These continuously duplicated medians would disturb the extraction of signal features for the motion recognition. Figure 5 shows a procedure for removing the duplicated median of signals. In the procedure, $\boldsymbol{Y}$ is the signal array of length N after filtering, and let $\boldsymbol{D} = \{d_0, d_1, \cdots, d_{(N-1)}\}$, which is an one-dimensional signal array used to store the signal values after removing the duplicated medians. The method of identifying the duplication is to check the equality of two successive signal values. If it is found, set **nil** to d_i; otherwise, set y_i to d_i. After processing the duplication, the motion signals, which can be used to extract the signal features, are displayed in Figure 4(b).

According to Table 1, the sets of feature's signals for a specific motion are located around peak values in the signal curve of Figure 4(b). To find out the

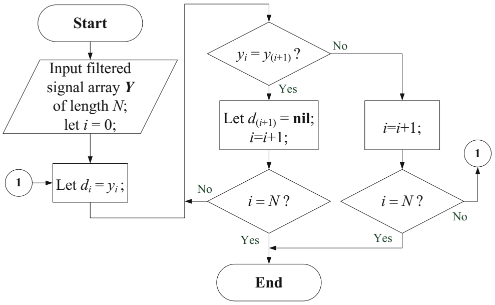

Start
Input filtered signal array $\boldsymbol{Y}$ of length N; let $i = 0$;
1
Let $d_i = y_i$;
$y_i = y_{(i+1)}$?
No
Yes
Let $d_{(i+1)}$ = **nil**; i=i+1;
i=i+1;
1
No
$i = N$?
$i = N$?
No
Yes
Yes
End

Fig. 5. The work flow of eliminating duplicated medians

features of motion's signals, the algorithm shown in Figure 6 is designed to locate a set of specific motion's features in the signal data. In the algorithm, let R be the sampling rate of the motion sensors, T the total signal acquisition time, *SwingNum* the number of swings to be acquired and analyzed, n the total number of acceleration's signals acquired in one typical swing, and p an adjustable parameter for *SwingNum*. In the feature searching algorithm mentioned above, the set of signals corresponding to one typical swing has to be identified prior to the computation of HP, BHP, and AHP. Through the calculation of the total signal acquisition time T and the number of swings *SwingNum* to get n for a typical swing, the three essential features, HP, BHP, and AHP, can be obtained. Subsequently, they are utilized to calculate the other five features listed in Table 1. In this research, the value of the adjustable parameter p is set to 5. One can lower p below 5, but not less than 2, to gain more features in the swing. The reason of not less than 2 is that in normal conditions, a complete swing cycle is not easy to be done under 2 seconds. In a circumstance in which one only wants to recognize the posture of subjects, one can set the p value above 5. Since the number of features collected is fewer, the recognition efficiency can be faster.

In addition, the position of the swinging arm is relevant to determine a motion posture. According to the way of attaching the sensors on the subject's arm, the acceleration values of the Y-axis display the arm's instant trajectory. As a swinging posture is in motion, we can extract the feature's values of HP, BHP, and AHP corresponding to the Y-axis from their known position on the X-axis. For example, to get a feature's value of HP on the Y-axis, one can refer to a value like x_5 on the X-axis and acquire the corresponding value on the Y-axis through the signal array $\boldsymbol{D}$. In Figure 7, since d_5, which is corresponding to x_5, is a feature of **nil** after duplicate's removal on X-axis, the next non-**nil** value d_6 should be used to correspond to x_5 for finding the value, i.e. y_6, on the Y-axis.

With the aforementioned schemes, we can extract 8 features from one measuring sensor. In the same way, we can also extract another 8 features from the second measuring sensor. Totally, we have 16 features that can be used in recognizing moving postures in our study. To find more accurate feature's borders, one needs to project the data of collected features through a kernel function

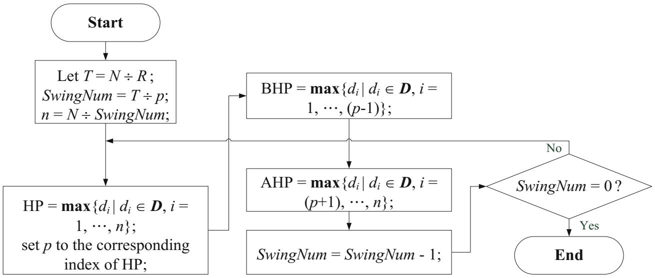

Start
Let $T = N \div R$;
$SwingNum = T \div p$;
$n = N \div SwingNum$;
BHP = **max**$\{d_i \mid d_i \in \boldsymbol{D}, i = 1, \cdots, (p\text{-}1)\}$;
No
HP = **max**$\{d_i \mid d_i \in \boldsymbol{D}, i = 1, \cdots, n\}$;
set p to the corresponding index of HP;
AHP = **max**$\{d_i \mid d_i \in \boldsymbol{D}, i = (p+1), \cdots, n\}$;
$SwingNum = 0$?
Yes
$SwingNum = SwingNum$ - 1;
End

Fig. 6. The work flow for feature searching

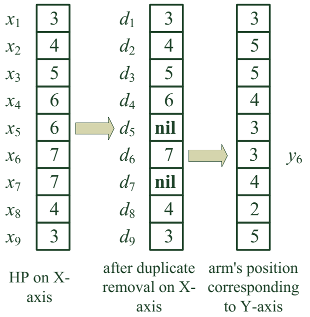


Fig. 7. Extraction of features on the Y-axis from the features on the X-axis

into a high dimension vector space for classification. We employ a radial basis function (Eq. (2)) as a mapping function to project the feature vectors onto the hyperspace. In Eq. (2), x is a feature's vector, x_i the mean vector of the feature's vectors mapped into the hyperspace, and σ the standard deviation of the projected vectors.

$$\phi(x) = (\|x - x_i\|) = \exp(-\frac{\|x - x_i\|^2}{\sigma}) \tag{2}$$

As training a SVM, we utilize a 10-fold cross validation scheme [17] [18] [19]. This is a practical validation method aimed at the generalization of the SVM outside the training data. The idea of this method is to separate the data into ten parts and to randomly select nine parts as a training reference and the remaining one as testing data. This training for the SVM takes a total of ten times. For each test, it shows appropriate accuracy. To evaluate the accuracy of the SVM, it is sufficient to take the average of the final accuracy results from the ten training.

In our investigation, the total recognition effectiveness of the trained SVM is examined by classifying the recognition samples into three sets, that is, the sets of truely and positively recognized postures (TPRP), falsely recognized as positive postures (FRPP), and falsely recognized as other postures (FROP).

The F-score (also F-measure) shown in Eq. (3) is used to evaluate the recognition effectiveness in the following experiments for each sport's category [20].

$$F = 2 \times \frac{Precision \times Recall}{Precision + Recall}, \tag{3}$$

where $Precision = \frac{\|TPRP\|}{\|TPRP\|+\|FRPP\|}$ and $Recall = \frac{\|TPRP\|}{\|TPRP\|+\|FROP\|}$.

4 Experimental Results

Two categories of experiments are demonstrated in this section. For each experiment, 10 subjects are recruited from the local campus. All of them enjoy normal health conditions. Prior to the experiments, the subjects were asked to fill in an information form about their personal information and experience of their sport as is illustrated in Table 2.

Table 2. Subjects' information

	Badminton	**Ping pong**	**Tennis**
Height distribution (cm)	153 ˜ 172	152 ˜ 170	153 ˜ 171
Weight distribution (Kg)	52 ˜ 78	40 ˜ 65	40 ˜ 67
Male	4	3	5
Female	6	7	5
Expert Level	1	1	0
Casual Level	1	1	2
Ordinary Level	8	8	8

4.1 Two Categories of Experiments

Experiment 1 – The *F*-score evaluation This experiment evaluates whether signal data collected from a single measuring device is sufficient to identify a single-handed swing. First, a trained SVM is utilized for recognizing the signal data from a single sensor attached on the front or upper arm. To compare the effect of using one sensor with that of two sensors, the F-score (Eq. (3)) is applied to demonstrate the effectiveness of swing posture recognition in different sport's category.

Experiment 2 – Testing the Accuracy of the Recognition Method This experiment tests the accuracy of the proposed method. That is, we build a complete motion feature database and utilize the trained SVM method to verify the accuracy of swing posture recognition for each sport category.

4.2 The *F*-Score Evaluation

Tables 3 and 4 display the F-score when using only one device. It can be seen from Table 3 that (as performing a single swing) the front arm's features cannot generate accurate feature's borders for the classification of sport's categories in the case of badminton and tennis. Although 86 correct samples are identified in ping pong, there are still 17 badminton samples and 78 tennis samples that are misread as ping pong samples. This significantly resulted in decreasing the F-score of the ping pong category. Obviously, the feature's signal data gathered from the front arm alone is not sufficient to identify postures.

Table 3. The F-score values when using only the front arm device

		Actual signals		
Recognition results		**Badminton**	**Ping pong**	**Tennis**
	Badminton	7	0	7
	Ping pong	17	86	78
	Tennis	76	14	15
	***F*-score**	0.12	0.61	0.15

Table 4. The F-score values when using only the upper arm device

		Actual signals		
Recognition results		**Badminton**	**Ping pong**	**Tennis**
	Badminton	83	3	5
	Ping pong	10	94	17
	Tennis	7	3	78
	***F*-score**	0.87	0.85	0.83

On the other hand, the sensor on the upper arm can detect a bigger movement signal as shown in Table 4. For swing posture recognition, The signal data of features with distinguished borders can be extracted. Consequently, the F-score with all sport's categories can achieve over 0.8. In comparison with the F-score in Table 3, we can see that the sensor position on the upper arm is much better than setting the sensor on the front arm. After combining the signal data detected from the two arm positions, we can acquire very good F-score as shown in Table 5. With the badminton and ping pong, the F-scores are both over 0.95; even for the tennis, which has lower recognition rate in the above experiments, its F-score can achieve to 0.94. Table 5 demonstrates that our proposed method keeps significant reliability.

4.3 Overall Recognition Rate

Table 6 demonstrates that the average recognition accuracy on the front and upper arms is 57% and 85%, respectively. With combining the two, the final

Table 5. F-score for the combination of the two feature's signal data

		Actual signals		
Recognition results		**Badminton**	**Ping pong**	**Tennis**
	Badminton	97	0	1
	Ping pong	2	98	7
	Tennis	1	2	92
	***F*-score**	0.98	0.95	0.94

Table 6. Swing posture recognition accuracy

	three front arm's features	**three upper arm's features**	**all eight features**
Badminton	7%	83%	97%
Ping pong	86%	94%	98%
Tennis	78%	78%	92%
Average	57%	85%	95.67%

average accuracy can reach up to 95.67%. In the badminton category, there are two data inputs recognized as ping pong, and one as tennis in Table 5. The reason is that when a subject swings too soft, the trained SVM may treat wrongly a badminton action as a ping pong action. If the arm position of the subject is too low, the trained SVM can misread badminton into tennis. In addition, two instances for ping pong is misread as tennis, which is caused by the subject's arm being too high. Regarding tennis, there is one instance in which the feature's signal data is misread as badminton, and seven as ping pong. Over powering in the ready and reaction phase of the sport, the subjects will cause a tennis action being read as badminton. Furthermore, tennis being read as ping pong is due to the result of hitting too hard in the ball hitting phase of the sport.

5 Conclusions

In this study, we propose a method of utilizing fewer triaxial accelerometers to recognize sport postures by attaching one triaxial accelerometer on the upper arm area and one on the front arm area. First, the swinging process that occurs in the three sports, badminton, tennis, and ping pong, are recognized. We then extract 8 features from the collected signal data, and use a trained SVM for recognition. The experimental results show that an accuracy of 95.67% is achieved by the proposed method. This research can be applied in training the athletes of badminton, tennis, and ping pong through adjustment in swinging postures to increase the effect of hitting, precision, and offensive play. In the future work, we try to develop other advanced methods with attaching appropriate accelerometers on subjects to collect more accurate data and recognize more complicated sports.

Acknowledgments. This work was supported by the National Science Council, Taiwan, under contracts NSC 99-2220-E-224-004. At the same time, the authors would like to thank those volunteers in the local campus, who ardently support the conduction of the sport experiments.

References

1. Chen, H.S., Chen, H.T., Chen, Y.W., Lee, S.Y.: Human action recognition using star skeleton. In: 4th ACM International Workshop on Video Surveillance and Sensor Networks, pp. 171–178. IEEE Press, CA (2006)
2. Xu, G., Huang, D., Zhou, Q., Zhao, D.: Modeling body motion posture recognition using 2d-skeleton angle feature. In: International Conference on Image, Vision and Computing (ICIVC 2012), Singapore (2012)
3. Rubaiyeat, H.A., Kim, T.S., Hasan, M.K.: Real-time recognition of daily human activities using a single tri-axial accelerometer. In: 5th International Conference on Embedded and Multimedia Computing (EMC), Cebu, pp. 1–5 (2010)
4. Reiss, A., Weber, M., Stricker, D.: Exploring and extending the boundaries of physical activity recognition. In: IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 46–50. IEEE Press, Anchorage (2011)
5. Zatsiorsky, V.M.: Kinematics of Human Motion. Technical report, Human Kinetics (1997)
6. Teixeiraa, T., Jung, D., Dublon, G., Savvides, A.: Recognizing activities from context and arm pose using finite state machines. In: Third ACM/IEEE International Conference on Distributed Smart Cameras, pp. 1–8. IEEE Press, Como (2009)
7. Zhu, C., Cheng, Q., Sheng, W.: Human activity recognition via motion and vision data fusion. In: Conference Record of the Forty Fourth Asilomar Conference on Signals, Systems and Computers (ASILOMAR), Pacific Grove, CA, pp. 332–336 (2010)
8. Jeong, K., Won, J., Bae, C.: User activity recognition and logging in distributed intelligent gadgets. In: IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems, pp. 683–686. IEEE Press, Seoul (2008)
9. Strohrmann, C., Harms, H., Kappeler-Setz, C., Trster, G.: Monitoring kinematic changes with fatigue in running using body-worn sensors. IEEE Transactions on Information Technology in Biomedicine 16, 983–990 (2012)
10. Karantonis, D.M., Narayanan, M.R., Mathie, M., Lovell, N.H., Celler, B.G.: Implementation of a real-time human movement classifier using a triaxial accelerometer for ambulatory monitoring. IEEE Transactions on Information Technology in Biomedicine 10, 156–167 (2006)
11. Thiemjarus, S.: A device-orientation independent method for activity recognition. In: International Conference on Body Sensor Networks (BSN), Singapore, pp. 19–23 (2010)
12. Stiefmeier, T., Roggen, D., Ogris, G., Lukowicz, P., Trster, G.: Wearable activity tracking in car manufacturing. IEEE Pervasive Computing 7, 42–50 (2008)
13. Stiefmeier, T., Lombriser, C., Roggen, D., Unker, B.J., Ogris, G., Troster, G.: Event-based activity tracking in work environments. In: 3rd International Forum on Applied Wearable Computing (IFAWC), Bremen, Germany, pp. 1–10 (2006)
14. Kao, T.P., Lin, C.W., Wang, J.S.: Development of a portable activity detector for daily activity recognition. In: IEEE International Symposium on Industrial Electronics, pp. 115–120. IEEE Press, Seoul (2009)

15. He, Z., Jin, L.: Activity recognition from acceleration data based on discrete consine transform and svm. In: IEEE International Conference on Systems, Man and Cybernetics, pp. 5041–5044. IEEE Press, San Antonio (2009)
16. Inoue, S., Hattori, Y.: Toward high-level activity recognition from accelerometers on mobile phones. In: International Conference on and 4th International Conference on Cyber, Dalian, pp. 225–231 (2011)
17. Courant, R., Hilbert, D.: Methods of Mathematical Physics. John Wiley & Sons (2008)
18. Kivinen, J., Smola, A.J., Williamson, R.C.: Online learning with kernels. IEEE Transactions on Signal Processing 52, 2165–2176 (2004)
19. Schaffer, C.: Technical note selecting a classification method by crossvalidation. Machine Learning 13, 135–143 (1993)
20. Gyllensten, I.C., Bonomi, A.G.: Identifying types of physical activity with a single accelerometer: Evaluating laboratory-trained algorithms in daily life. IEEE Transactions on Biomedical Engineering 58, 2656–2663 (2011)

Information Hiding Based on Binary Encoding Methods and Linear Transformation Scrambling Techniques

Kuang Tsan Lin

Department of Mechanical and Computer-Aided Engineering, St. John's University
499, Section 4, Tam King Road, Tamsui, New Taipei City 25135, Taiwan
ktlin@mail.sju.edu.tw

Abstract. The paper proposes a hybrid method that combines a binary encoding method and a linear transformation scrambling technique to hide an image. Firstly, a linear transformation scrambling technique is used to rearrange the pixel values of a covert image to form a linear-transformation-scrambled matrix by using a certain linear transformation scrambling technique. Secondly, the linear-transformation-scrambled matrix is encoded into a host image to form an overt image by using a certain encoding method. The overt image contains seven groups of binary codes, i.e. identification codes, dimension codes, graylevel codes, linear slope codes, linear intersection codes, linear transformation scrambling times codes, and information codes. The parameters are used to encode and hide the covert image. According to the simulation results, the proposed method does well, larger image scrambling degree for the scrambled matrix, and saver computing time.

Keywords: Information hiding, Binary encoding, Linear transformation scrambling technique.

1 Introduction

Information scrambling techniques have been widely used in the information hiding and to enhance its information security. Some studies have proposed different approaches of image scrambling techniques. Using Arnold transformations [1] and p-Fibonacci transformations [2] can do well but lesser image scrambling degrees for lesser scrambling times and unstable values of the image scrambling degree for the scrambling times process. Using crossover mechanism of genetic algorithms [3] have large values of the image scrambling degree for the scrambling times process but time-consuming.

Some other studies have proposed to assemble image encoding methods and information scrambling techniques to enhance information security. Zhao et al. [4] assembled a fractional Fourier transform method and an image scrambling technique to conceal an image. Meng et al. [5] combined an iterative Fresnel transform method and an image scrambling technique to hide an image. They have very good robustness for the hybrid methods, but they might have some distortion in reconstructed images.

The paper presents an information hiding method for images based on a binary encoding method [6] and a linear transformation scrambling technique to hide a covert

H. Sun et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 329, DOI: 10.1007/978-3-319-12286-1_17

image in a host image to form an overt image. The hybrid method can be applied to equilateral and non-equilateral images, and it is difficult to distinguish between the overt image and the corresponding host image. And, the hybrid method has a good security, and larger image scrambling degree for the lesser scrambling times.

2 Linear Transformation Scrambling Techniques for Encoding Images

Let $\boldsymbol{C}$ be a $r\times c$ covert image to be linear-transformation-scrambled and let $\boldsymbol{D}$ be a $r\times c$ matrix formed from the linear transformation scrambling technique of $\boldsymbol{C}$. The processes for deriving $\boldsymbol{D}$ from $\boldsymbol{C}$ are shown below. First, transform the pixels of $\boldsymbol{C}$ into form a pixel string $\boldsymbol{A}$ with $r\times c$ elements according to a specified order (from the first row to the last row and from the first column to the last column for the same row). Second, assign a linear slope m, a linear intersection b, and transform the pixel string $\boldsymbol{A}(u)$ with $r\times c$ elements to a scrambled pixel string $\boldsymbol{B}(v)$ with $r\times c$ elements by $v=mod((m\times(u\text{-}1)+b)/(r\times c))$. Third, determine the linear-transformation-scrambled matrix $\boldsymbol{D}$ with $r\times c$ by using a specified order (every c elements to form a row from the first element to the last element) from the scrambled pixel string $\boldsymbol{B}(v)$ with $r\times c$ elements. An example for the linear transformation scrambling processes with linear slope and intersection equal to 3 and 5, respectively, from a 4×4 covert image $\boldsymbol{C}$ to a 4×4 scrambled matrix $\boldsymbol{D}$ with a linear slope m=3, a linear intersection b=5, and is depicted in Fig. 1.

$$\boldsymbol{C}=\begin{bmatrix}1 & 2 & 3 & 4\\ 5 & 6 & 7 & 8\\ 9 & 10 & 11 & 12\\ 13 & 14 & 15 & 16\end{bmatrix}$$

(a)

$$\boldsymbol{A}=\begin{bmatrix}1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 & 16\end{bmatrix}$$

(b)

$$\boldsymbol{B}=\begin{bmatrix}6 & 9 & 12 & 15 & 2 & 5 & 8 & 11 & 14 & 1 & 4 & 7 & 10 & 13 & 16 & 3\end{bmatrix}$$

(c)

$$\boldsymbol{D}=\begin{bmatrix}6 & 9 & 12 & 15\\ 2 & 5 & 8 & 11\\ 14 & 1 & 4 & 7\\ 10 & 13 & 16 & 3\end{bmatrix}$$

(d)

Fig. 1. Example for linear transformation scrambling processes from $\boldsymbol{C}$ to $\boldsymbol{D}$. (a) assumed covert matrix $\boldsymbol{C}$; (b) the pixel string $\boldsymbol{A}$; (c) the scrambled pixel string $\boldsymbol{B}$; (d) scrambled matrix $\boldsymbol{D}$ with m=5 and b=3

3 Binary Encoding Methods for Hiding Linear-Transformation-Scrambled Matrices

Assume ***H*** is an $M \times N$ host image used to hide a linear-transformation-scrambled matrix ***D*** to form an $M \times N$ overt image ***H****, where the linear-transformation-scrambled matrix ***D*** is processed a linear transformation scrambling technique from a covert image ***C***. All of the pixels in ***H**** are classified into seven groups, i.e. identification codes, dimension codes, graylevel codes, linear slope codes, linear intersection codes, linear transformation scrambling times codes, and information codes.

The identification codes are used to justice whether the codes in ***H**** is encoded with the proposed encoding method or not; the dimension codes are used to denote the dimensions of the covert image ***C***; the graylevel codes are used to denote the graylevel of the covert image ***C***; the linear slope codes are used to denote the slope of the linear equation; the linear intersection codes are used to denote the intersection of the linear equation; the linear transformation scrambling times codes are used to denote the times of the linear transformation scrambling technique repeated. The information codes are used to hide the linear-transformation-scrambled matrix ***D*** and they are encoded at the second row to the last row of ***H****. Apart from this, the other seven groups of binary codes are encoded at the first row of ***H**** with the orders specified by the designer.

For the identification codes, the number of the codes must be large enough to avoid incorrect judgment and the codes are binary, e.g. 11000110001110011110011110111101.

For the dimension codes, they are two sets of ten-digit binary codes. The first set of binary codes is used to denote the row dimension r ($r > 1$) of ***C*** and it includes $r1$ to $r10$. The relationship between r and $r1$- $r10$ is followed by Eq. (1). The second set of binary codes is used to denote the column dimension c ($c > 1$) of ***C*** and it includes $c1$ to $c10$. The relationship between c and $c1$- $c10$ is followed by Eq. (2).

$$r = \sum_{i=1}^{10} r_i \cdot 2^{i-1} + 2\,, \tag{1}$$

$$c = \sum_{i=1}^{10} c_i \cdot 2^{i-1} + 2\,. \tag{2}$$

For the graylevel codes, they are eight-digit binary codes $g1$ to $g8$, and they are used to denote the g value of ***C***. The relationship between g and $g1$- $g8$ is similar to Eq. (3).

$$g = \sum_{i=1}^{8} g_i \cdot 2^{i-1} + 1\,. \tag{3}$$

The linear slope codes, they are eight-digit binary codes $m1$ to $m8$, and they are used to denote the linear slope m of the linear transformation. The relationship between m and $m1$- $m8$ is similar to Eq. (4).

$$m = \sum_{i=1}^{8} m_i \cdot 2^{i-1} + 1 . \tag{4}$$

The linear intersection codes, they are eight-digit binary codes $b1$ to $b8$, and they are used to denote the linear intersection b of the linear transformation. The relationship between b and $b1$- $b8$ is similar to Eq. (5).

$$b = \sum_{i=1}^{8} b_i \cdot 2^{i-1} + 1 . \tag{5}$$

For the linear transformation scrambling times codes, they are ten-digit binary codes $t1$ to $t10$, and they are used to denote the t times of the linear transformation scrambling technique repeated. The relationship between t and $t1$- $t10$ is similar to Eq. (6).

$$t = \sum_{i=1}^{10} t_i \cdot 2^{i-1} . \tag{6}$$

For the information codes, they are used to encode ***D*** and the encoding processes to encode ***D*** in the $M \times N$ host image H to form an $M \times N$ overt image ***H**** are brief stated below.

(1) Create a binary array ***R*** with N elements. Some of the elements of ***R*** contain identification codes, dimension codes, graylevel codes, linear slope codes, linear intersection codes, and linear transformation scrambling times. The other elements of R are not used and they are all set to be 0.

(2) Transform the elements of ***D*** to form the elements of a linear-transformation-scrambled string ***E*** followed by

$$\boldsymbol{E}((r'-1) \times c + c') = \boldsymbol{D}(r', c') , \tag{7}$$

where $1 \le r' \le r$ and $1 \le c' \le c$.

(3) After $\boldsymbol{E}(k) = \sum_{i=0}^{g-1} a_i(k) \cdot 2^{i-1}$ are known, transform ***E*** into a binary data string ***F*** according to

$$\boldsymbol{F}(i + h \times (k-1)) = a_i(k) . \tag{8}$$

(4) Use the elements of ***F*** to form the elements of a $(M\text{-}1) \times N$ data matrix ***S***. Since the array data number L of ***F*** may be smaller than $(M\text{-}1) \times N$, there are $(M\text{-}1) \times N\text{–}L$ dummy elements in ***S*** not formed from the elements of ***F***. As a result, the values of dummy elements are all set to be 0.

(5) Combine the row array ***R*** with N elements and the $(M\text{-}1) \times N$ matrix ***S*** to form a $M \times N$ binary matrix ***T***. The first row of ***T*** is duplicated from ***R***, while other rows of T are duplicated from ***S*** in succession.

(6) Modify the element ***H***(u,v) of the host image ***H*** to form the element $\boldsymbol{H}'$ (u,v) of a modified matrix $\boldsymbol{H}'$ followed by

$$\boldsymbol{H}'(u,v)=2\times floor(\boldsymbol{H}(u,v)/2), \tag{9}$$

where the function $floor(x)$ modulates the value of x to the nearest integer x_n ($< x$), and every $\boldsymbol{H}'(u,v)$ is an even integer.

(7) An overt image $\boldsymbol{H}^*$ is formed by summing the corresponding elements of matrices $\boldsymbol{T}$ and $\boldsymbol{H}'$, i.e.

$$\boldsymbol{H}^*(u,v)=\boldsymbol{T}(u,v)+\boldsymbol{H}'(u,v)\,. \tag{10}$$

Figure 2 depicts an assumed host matrix $\boldsymbol{H}$ and an assumed binary matrix $\boldsymbol{T}$, and the resulted modulated matrix $\boldsymbol{H}'$ and the resulted overt matrix $\boldsymbol{H}^*$.

$$\boldsymbol{H}=\begin{bmatrix} 9 & 10 & 11 \\ 1 & 0 & 1 \\ 77 & 75 & 73 \end{bmatrix} \quad \boldsymbol{T}=\begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \quad \boldsymbol{H}'=\begin{bmatrix} 8 & 10 & 10 \\ 0 & 0 & 0 \\ 76 & 74 & 72 \end{bmatrix} \quad \boldsymbol{H}^*=\begin{bmatrix} 9 & 10 & 11 \\ 0 & 1 & 0 \\ 77 & 75 & 73 \end{bmatrix}$$

(a) (b) (c) (d)

Fig. 2. (a) assumed host matrix $\boldsymbol{H}$; (b) assumed binary matrix $\boldsymbol{T}$; (c) resulted modulated matrix $\boldsymbol{H}'$; (f) resulted overt image $\boldsymbol{H}^*$

For a $r\times c$ image $\boldsymbol{D}$ scrambled by the crossover mechanism of genetic algorithm from a $r\times c$ image $\boldsymbol{C}$, the definition of the image scrambling degree δ is [7]

$$\delta=\frac{\sum_{i=1}^{r}\sum_{j=1}^{c}[\boldsymbol{W}_{-1,0}(i,j)+\boldsymbol{W}_{1,0}(i,j)+\boldsymbol{W}_{0,-1}(i,j)+\boldsymbol{W}_{0,1}(i,j)]}{255^2\times r\times c}, \tag{11a}$$

where

$$\boldsymbol{W}_{m,n}=\left|\,[\boldsymbol{D}(i+m,j+n)-\boldsymbol{D}(i,j)]^2-[\boldsymbol{D}(i+m,j+n)-\boldsymbol{D}(i,j)]^2\right|. \tag{11b}$$

A larger δ value indicates that $\boldsymbol{C}$ and $\boldsymbol{D}$ are more different.

The *PSNR* values of the two images $\boldsymbol{H}$ and $\boldsymbol{H}^*$ is used to check image quality. The definition of *PSNR* is [8]

$$PSNR=10\times\log\left(\frac{M\times N}{MSE}\right), \tag{12a}$$

where

$$MSE=\frac{1}{M\times N}\sum_{i=1}^{M}\sum_{j=1}^{N}[\boldsymbol{H}(i,j)-\boldsymbol{H}^*(i,j)]^2\,. \tag{12b}$$

Basically, if the *PSNR* is larger than 30, it will be difficult to distinguish the difference between $\boldsymbol{H}$ and $\boldsymbol{H}^*$ with naked eyes; that is the image $\boldsymbol{H}^*$ looks nearly the same as $\boldsymbol{H}$ [9].

4 Simulations

Figure 3 shows a 64 × 64 256-graylevel image as the covert image ***C***. Fig. 4 shows a 256 × 256 256-graylevel image used as the host image ***H***.

The simulation about the covert image ***C*** in Fig. 3 is introduced below. Because the dimension of ***H*** is 256 × 256, the dimension of the row array ***R*** is 1 × 256. The 1st to 32nd elements of ***R*** are used to be the identification codes, and they are designated as 11000110001110011110011110111101. The 33rd to 52nd elements of ***R*** are used to be the covert image dimension codes. Since the size of ***C*** is 64 × 64, the two sets of dimension codes for *r* and *c* are 0000111110. The 53rd to 60th elements of ***R*** are used to be the graylevel codes. Since the number of the gray values is 256 (=2^8), *g*=8, the codes are 00000111. The 61st to 68th elements of ***R*** are used to the linear slope codes. Since *m* is set to be 71 here, the codes are 01000110. The 69th to 76th elements of ***R*** are used to the linear intersection codes. Since *b* is set to be 3 here, the codes are 00000010. The 77th to 86th elements of ***R*** are used to the linear transformation scrambling times codes *t*. Since *t* is set to be 2 here, the codes are 0000000010.

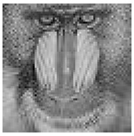


Fig. 3. A covert image for test

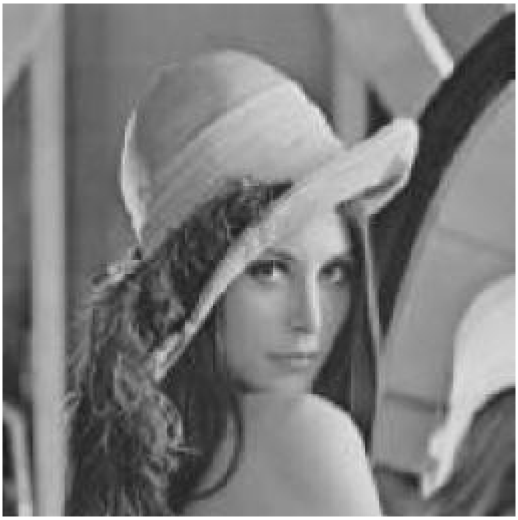
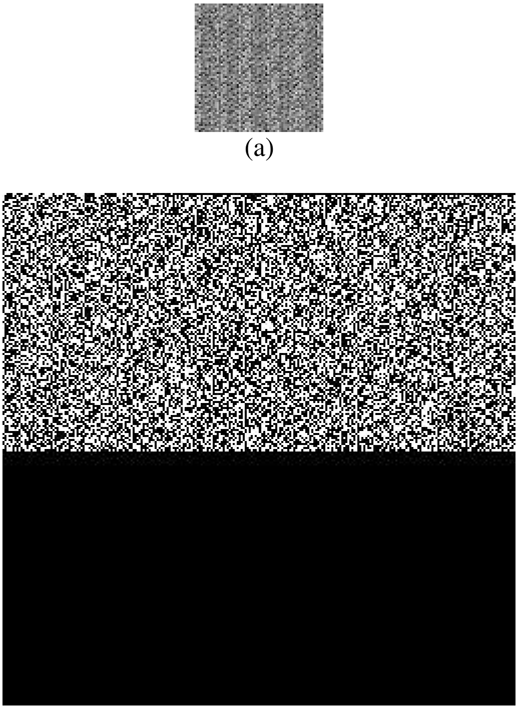


Fig. 4. A host image for encoding to form an overt image

First we scramble the 64×64 covert image $\boldsymbol{C}$ into a 64×64 matrix $\boldsymbol{D}$ by using the proposed linear transformation scramble technique. The linear-transformation-scrambled matrix $\boldsymbol{D}$ is depicted in Fig. 5(a). Then we transform $\boldsymbol{D}$ into a linear-transformation-scrambled string $\boldsymbol{E}$ with 4096 elements. Then, transform $\boldsymbol{E}$ into a binary-data string $\boldsymbol{F}$ with 32768 elements. Subsequently, copy the elements of $\boldsymbol{F}$ to form the elements of a 255×256 data matrix $\boldsymbol{S}$. The 32769th to 65280th elements of $\boldsymbol{F}$ are all set to be 0.

Second, we change $\boldsymbol{F}$ into a 255×256 matrix $\boldsymbol{S}$, and we combine the 1×256 binary row array $\boldsymbol{R}$ and the 255×256 matrix $\boldsymbol{S}$ to form a 256×256 binary matrix $\boldsymbol{T}$. The matrix $\boldsymbol{T}$ is depicted in Fig. 5(b). Furthermore, the host image $\boldsymbol{H}$ is modulated to form a modified image $\boldsymbol{H}'$. Then, the corresponding elements of the matrices $\boldsymbol{T}$ and $\boldsymbol{H}'$ are summed to form an overt image $\boldsymbol{H}^*$. The modified image $\boldsymbol{H}'$ and overt image $\boldsymbol{H}^*$ look nearly the same as the host image $\boldsymbol{H}$ in Fig. 4.

The *PSNR* value between $\boldsymbol{H}$ and $\boldsymbol{H}^*$ is equal to 51.1 for the case of the covert binary image. Therefore, the two images $\boldsymbol{H}$ and $\boldsymbol{H}^*$ look almost identical for both the case. Moreover, for the case the *PSNR* value between the original covert image $\boldsymbol{C}$ and the decoded covert image $\boldsymbol{C}^*$ is infinity, i.e. there is no distortion during the covert image decoding.



(a)

(b)

Fig. 5. (a) The linear-transformation-scrambled matrix $\boldsymbol{D}$; (b) matrix $\boldsymbol{T}$; (c) the overt matrix $\boldsymbol{H}^*$ for encoding the covert image in Fig. 3

5 Discussions

The proposed combine the binary encoding method and the linear transformation scrambling technique are demonstrated only for brief and clear in this paper, but rearranging the orders of codes or rearranging pixel positions for encoding codes can get a better security for the proposed method.

We compute the image scrambling degree of the scrambled images in Fig. 3 for different algorithm. In Table 1 shows some results. Firstly, bigger *a* (linear slope) can get bigger the image scrambling degree δ by the proposed method, and *b* (linear intersection) is independent of δ. Secondly, δ by using the proposed method is mostly larger than δ by using the Arnold and p-Fibonacci transformation for the scrambling time *t*=1 or 2, and the proposed method is better security than other transformations. In Fig. 6 shows δ corresponding to the covert image in Fig. 3 by the proposed method, δ are not stable for different *t*. And the proposed method is much saver time than the Lin and Lin's paper [3].

Table 1. The image scrambling degree of the scrambled image for different algorithms

t	The proposed method						Arnold method	p-Fibonacci method
	a=7	*a*=71	*a*=171	*a*=271	*a*=371	*a*=471		
1	0.1402	0.1415	0.1702	0.1672	0.1684	0.1796	0.1249	0.1514
2	0.1649	0.1715	0.1578	0.1725	0.1744	0.1818	0.1661	0.1607

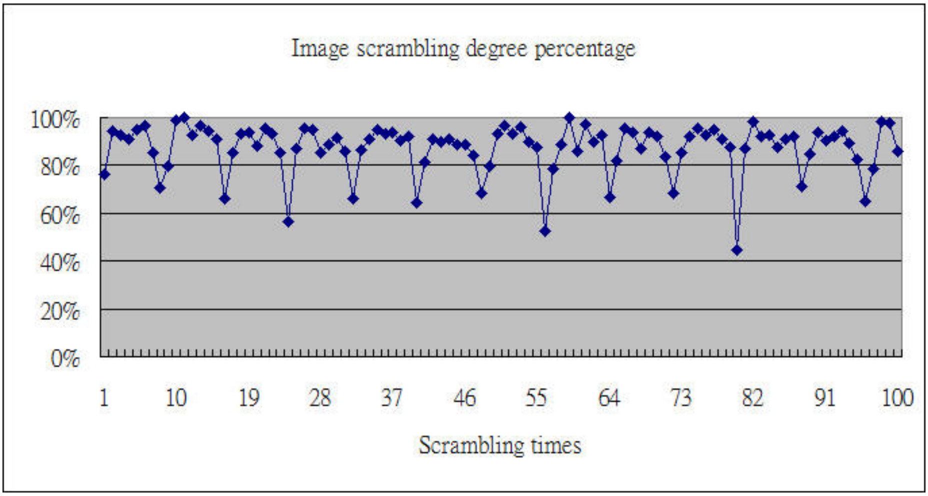


Fig. 6. Image scrambling degree percentages with different *t* values and by using the proposed method

6 Conclusions

The proposed method combines the binary encoding and the linear transformation scrambling technique to encode an image. The linear transformation scrambling technique scrambles into a linear-transformation-scrambled matrix ***D*** from the covert image ***C***. The linear-transformation-scrambled matrix ***D*** can obtain a bigger value of the image scrambling degree and get a saver computing time. The overt image ***H**** hides the linear-transformation-scrambled matrix ***D*** by using the binary encoding method. Thanks to the *PSNR* of ***H**** and ***H*** is larger than 50, the two images ***H*** and ***H**** look nearly the same. And, the overt image ***H**** can directly recover the decoded covert image ***C**** and not need the host image ***H***. Apart from these, thanks to the *PSNR* of ***C*** and ***C**** is infinity, ***C*** is the same as ***C****, i.e. it has not any distortion for the decoding processes of the covert image.

References

1. Yang, Y.L., Cai, N., Ni, G.Q.: Digital Image Scrambling Technology Based on the Symmetry of Arnold Transform. Journal of Beijing Institute of Technology 15, 216–221 (2006)
2. Zhou, Y., Agaian, S., Joyner, V.M., Panetta, K.: Two Fibonacci p-code based image scrambling algorithms. In: Proceedings of SPIE, vol. 6812, p. 681215 (2008)
3. Lin, K.T., Lin, P.H.: Information hiding based on binary encoding methods and crossover mechanism of genetic algorithms. In: Pan, J.-S., Krömer, P., Snášel, V. (eds.) Genetic and Evolutionary Computing. AISC, vol. 238, pp. 203–212. Springer, Heidelberg (2014)
4. Zhao, J., Lu, H., Fan, Q.: Color Image Encryption Based on Fractional Fourier Transforms and Pixel Scrambling Technique. In: Proceedings of SPIE, vol. 6279, p. 62793B (2007)
5. Meng, X.F., Cai, L.Z., Yang, X.L., Shen, X.X., Dong, G.Y.: Information Security System by Iterative Multiple-phase Retrieval and Pixel Random Permutation. Applied Optics 45, 3289–3295 (2006)
6. Lin, K.T.: Digital Information Encrypted in an Image Using Binary Encoding. Optics Communications 281, 3447–3453 (2008)
7. Yu, X.Y., Ren, H., Li, E.S., Zhang, X.D.: A New Measurement Method of Image Encryption. Journal of Physics: Conference Series 48, 408–413 (2006)
8. Kutter, M., Petitcolas, F.A.P.: A Fair Benchmark for Image Watermarking Systems. In: Proceedings of SPIE, vol. 3657, pp. 226–231 (1999)
9. Shih, T.K., Lu, L.C., Chang, R.C.: An Automatic Image in Paint Tool. In: Proceedings of the Eleventh ACM International Conference on Multimedia, pp. 102–103 (2003)

Dispersed Data Hiding Using Hamming Code with Recovery Capability

Brian K. Lien*, Shan-Kang Chen, Wei-Sheng Wang, and Kuan-Pang King

Dept. of Computer Science and Information Engineering, Fu Jen Catholic Univ., Taiwan
002954@mail.fju.edu.tw

Abstract. Hamming codes can improve the embedding efficiency by hiding messages in a block-by-block manner with pixel-flipping. But since each pixel in the block is not dispersed in the image, it can only be flipped individually thus introducing undesirable visual distortion for halftone images. Also, since the errors caused by tamper are usually more than one bit in a block, the tampered region cannot be recovered by error correction of Hamming code. This paper proposes a dispersed block generating scheme through Space-filling curve decomposition to hide data using Hamming coding into these dispersed blocks. Each block consists of pixels randomly and uniformly distributed all over the cover halftone image, and the relation between pixels in the adjacent blocks is the adjacent pixels along the Space-filling curve. Experimental results show that the proposed method significantly improves the visual quality of marked halftone images and can recover local tamper.

Keywords: Data hiding, Halftone, Space-filling Curve, Hamming code.

1 Introduction

Digital halftoning, which converts multi-tone images into two-tone format, is a critical technique in printing process. There are many halftone methods, and the most popular ones are the ordered dithering and error diffusion. Ordered dithering compares the pixels in the original graylevel image with a periodic and deterministic threshold matrix and error diffusion employs neighborhood operations to reduce the quantization error [1].

The increasingly prevalent usage of network and multimedia has heightened the need for exchanging digital documents in printed format, thus it has become necessary to hide data in the halftone images for the purpose of copyright protection or content authentication. In general, the data hiding methods for halftone images can be divided into two categories. The methods in the first category embed data during the halftoning process, having good visual quality, but requiring the original grayscale image. Whereas the methods in the second category embed data directly into the halftone images after they have been generated. The advantage of this kind of methods is that they can be applied to hide information for all kinds of halftone images.

* Corresponding author.

H. Sun et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 329, DOI: 10.1007/978-3-319-12286-1_18

Block-wise data hiding scheme is commonly adapted to hide data directly into the cover image. It can be applied as fragile watermarking methods [2, 3] to verify the integrity of the received images or as robust methods with codeword of larger Hamming distance [4-8]. It can also be used to increase the visual quality by selecting a best location in the block according to some scores [9], or to achieve reversibility [10-14].

Instead of hiding only one bit into a block of pixels, high embedding efficiency data hiding schemes using covering codes were proposed to embed multiple bits into a block by forcing the block of pixels to have certain relation [15, 16]. A method called Data Hiding using Hamming Code (DHHC) was proposed recently [17] to hide data into halftone images. Specifically, given a 2^p-1 image block, the scheme can embed p bits of data by flipping at most a pixel. But for halftone images, changing individual pixels will introduce undesirable visual distortion. A proved better embedding technique to improve visual quality for halftone images is to exchange neighboring pixels instead of flipping individual pixels [18]. In [19], a new data hiding scheme was proposed to improve [17] by changing pixels in pairs rather than individually. However, it sacrificed half of image pixels for pair toggling, thus the data hiding capacity reduced by half.

DHHC divided the cover image into continuous blocks of size 4×4, since each pixel in the block is clustered, it could only flip individual pixel thus introducing undesirable visual distortion. The goal of image authentication is to verify that an image has not been altered during communication. DHHC could locate the tampered region by checking the corresponding authentication mark. However since the errors caused by tamper are usually more than one bit in a block, the tampered region cannot be recovered.

This paper proposes a dispersed block generation scheme through Space-filling curve decomposition to hide data using Hamming coding into these dispersed blocks. Each block consists of pixels randomly and uniformly distributed all over the cover halftone image, and the relation between pixels in the adjacent blocks is the adjacent pixels along the Space-filling curve. The proposed scheme decreases the chance of forced single-pixel toggling by a forward search which selects the best pair of neighboring pixels along Space-filling curve for a to-be-embedded bit. For a local tamper, only a pixel in a block will be altered. Thus we can apply the Hamming coding to recover it. The method is named as Dispersed Data Hiding using Hamming Code (DDHHC). Experimental results show that the proposed method significantly improves the visual quality of marked halftone images and can recover local tamper.

The remainder of this paper is organized as follows. In Section 2, we describe the dispersed block generating scheme associated with a forward search first, then, detail an implementation of data hiding using Hamming code based on the proposed scheme. Section 3 presents our experimental results. Finally, we conclude in Section 4.

2 Dispersed Data Hiding Using Hamming Code

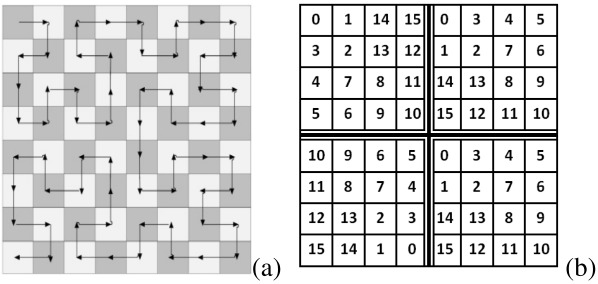
2.1 Data Hiding Using Hamming Code (DHHC)

DHHC employed hamming code (15, 11) to hide data for a halftone image which hided 4-bits in a 4×4 block by flipping at most a pixel. In DHHC, the cover image was directly divided into blocks of size 4×4. 15 bits from a block is regarded as a code word. The syndrome computed from the codeword and a 4-bit secret message is XORed to conceal this secret message in the codeword. If the result of a XOR operation is not zero, the corresponding position in the codeword should be flipped.

DHHC divided the cover image into continuous blocks of size 4×4, since each pixel in the block is not dispersed in the halftone image, the pixel to be changed to its opposite value can only be flipped individually thus introducing undesirable visual distortion. The goal of image authentication is to verify that an image has not been altered during communication. DHHC located the tampered region by checking the corresponding authentication mark. However since the errors caused by tamper are usually more than one bit in a block, the tampered region cannot be recovered.

2.2 The Space-Filling Curve Block Generating Scheme

A space-filling curve is a continuous mapping of a closed unit interval into a closed unit square. The existence was first discovered by Peano in 1890. Hilbert was the first to describe a procedure to construct such a curve. The Hilbert space filling curve visits every point in a square grid with a size of power of 2. For example, Fig. 1(a) shows a Hilbert curve of size 8x8. This curve preserves the locality in the multidimensional space. A proof showing that the Hilbert space-filling curve achieves the best clustering among the exiting space filling curves was given by Moon et al. [20]. Because of this locality-preserving property, it has many applications in a variety of areas.



0	1	14	15	0	3	4	5
3	2	13	12	1	2	7	6
4	7	8	11	14	13	8	9
5	6	9	10	15	12	11	10
10	9	6	5	0	3	4	5
11	8	7	4	1	2	7	6
12	13	2	3	14	13	8	9
15	14	1	0	15	12	11	10

Fig. 1. An example of Hilbert curve decomposition for an 8x8 region

To decompose an image into N blocks of M pixels, the image is first divided into M subimages. Each subimage has N pixels. Then, the sequence of pixels along a space-filling curve on each subimage become a linear lists; denoted as L_i, i = 0,1,…N-1, each list contains N connected locations. Those pixels corresponding to the same index k in each list are gathered as a block of M pixels, denoted as B_k. As an example, a small halftone image of size 8x8 is divided into 4 subimages and each list in the subimage is formed by replicating the one-dimensional array of 0, 1... 15 along the Hilbert curve as shown in Fig. 1(b). As you can see in Fig. 1(b), pixels in each block are randomly dispersed in the halftone image. Each block contains 4 pixels. The i-th element of B_k and $B_{(k+1)}$ are the horizontal or vertical neighbors along the Hilbert curve.

Multiple bits can be embedded in each block by forcing its M pixels to have certain relation according to the data bits to be embedded. To satisfy the relation, certain pixels in the block may need to be changed to its opposite value. Instead of directly flipping a pixel, we toggle it by a forward search based on a selection criterion. Some existing criteria can be found in [9] [18]. If a pixel in a block needs to be changed to its opposite value, the forward search will search for a best pair of pixels to toggle.

Let's observe two examples of forward search on Hilbert curve and raster scan covered by a 3×3 window respectively as shown in Fig. 2. The shadow cells represent the pixels that have been visited. The center pixel is the pixel to be flipped by swapping one of its unvisited neighbors. As can be seen from Fig. 2, the possible flappable neighbors for a selected location are those unvisited neighboring pixels along the Space-Filling curve. For example, there are at least 4 unvisited neighbors for the raster scan space-filling curve except the boundary pixels.

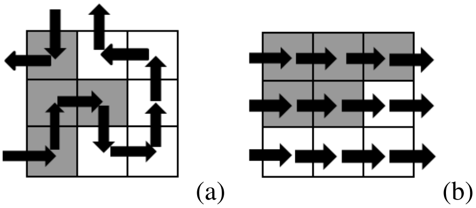


Fig. 2. Two examples of forward search on (a) Hilbert curve and (b) raster scan respectively covered by a 3×3 window

2.3 The Proposed Method

In this section we apply the scheme described in the previous section to improve DHHC. The proposed method is named as Dispersed Data Hiding Hamming Code (DDHHC). First, we divide the image into 16 subimages as shown in Fig. 3.

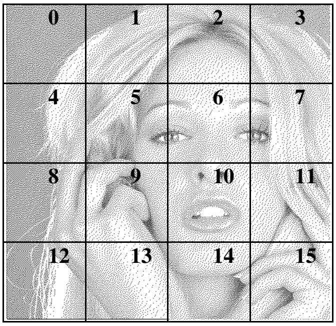


Fig. 3. The subimages of DDHHC partition

Second, we index the pixels of each subimage according to the sequence number of the space-filling curve as described in the previous section. Those pixels corresponding to the same index in each subimage are gathered as a block of 16 pixels. For a halftone image of size 512×512, there are total 16384 blocks, thus the embedding capacity is 65536 bits. Each block is indexed using the same sequence number of the space-filling curve.

To embed 4×m bits of secret message, we randomly select m blocks from 16384 blocks. Then these m blocks are sorted by the index number and each 4-bits secret message is embedded into the 15-bits codeword randomly selected from the corresponding ordered block in a block-by-block manner.

Unlike DHHC flipping the value in the selected location, DDHHC pair-toggles it with one of its neighboring pixels if possible to avoid the disturbance of local average intensity. Using the forward search as describe in the previous section with connection selection criteria of [18], the best toggled pixel pair will be the one which has the minimum connection value after toggling.

To read the embedded data, the marked image is partitioned into blocks using the same Space-Filling curve partition. Then the embedded data bits can be extracted by examining the syndrome computed from the 15-bits codeword of each block.

3 Experimental Result

In this section, we conduct a number of experiments to show the performance of the proposed technique. The tested images are 512×512 Floyd error diffusion halftone images Lena, Barbara, Tiffany, Pepper and Gold hill. The capacity is 65,536 bits with at most 16,384 bits toggling.

3.1 Visual Quality

First, we compare the number of self-toggling incurred. Table I shows the average number of self-toggling incurred when embedding 4096, 16,384 and 65,536 bits into the tested 512×512 images respectively. As can be seen from these Table I, the number of self-toggling is dramatically reduced by our method in all the tested images.

Table 1. The comparison of the number of self-toggling

Embedded bits	65536		16384		4096	
	DHHC	DDHHC	DHHC	DDHHC	DHHC	DDHHC
Lena	15485	2143	3866	41	964	2
Barbara	15412	2420	3851	66	962	6
Tiffany	15500	6066	3874	823	967	177
Pepper	15488	2647	3867	161	968	32
Gold hill	15394	2627	3849	113	961	15
Average	15456	3181	3862	241	964	46

In the perceptual quality comparison, we adopt the Modified Peak Signal to Noise Ratio (MPSNR) to measure the visual quality of the marked halftone images. Here, the MPSNR is defined as the PSNR between the filtered original halftone image and the filtered marked halftone image with a lowpass filter simulating human visual system. In Table II, we show the average MPSNR comparison of embedding 4096, 16384 and 65,536 bits into the tested 512×512 images respectively. The result shows that the proposed method achieves much better quality than DHHC.

Table 2. MPSNR comparison in marked error diffusion Halftone Image

Embedded bits	65536		16384		4096	
	DHHC	DDHHC	DHHC	DDHHC	DHHC	DDHHC
Lena	25.71	30.57	32.03	38.60	38.12	44.71
Barbara	25.69	30.38	32.03	38.57	38.14	44.77
Tiffany	24.71	27.15	31.72	36.24	38.04	42.91
Pepper	25.60	29.90	31.98	38.02	38.1	44.19
Gold hill	25.66	30.30	32.02	38.66	38.13	44.96
Average	25.47	29.66	31.96	38.02	38.11	44.31

Finally, Fig 4, 5 and 6 shows the marked halftone images for subjective quality evaluation, where image (a) is processed by DHHC and (b) is by DDHHC method. From these images, we can see DDHHC has much less undesirable visual distortion than DHHC.

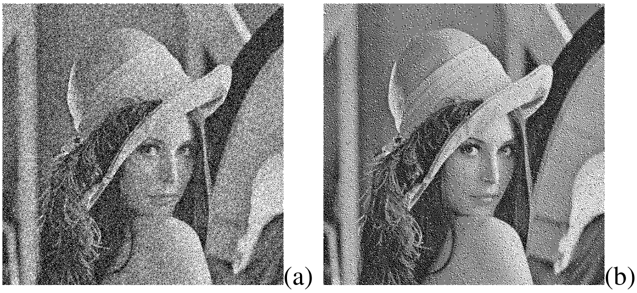


Fig. 4. The 65536 bits marked error diffusion Lena

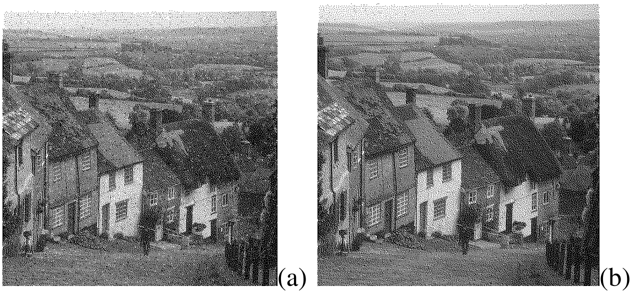


Fig. 5. The 16384 bits marked error diffusion Gold-hill

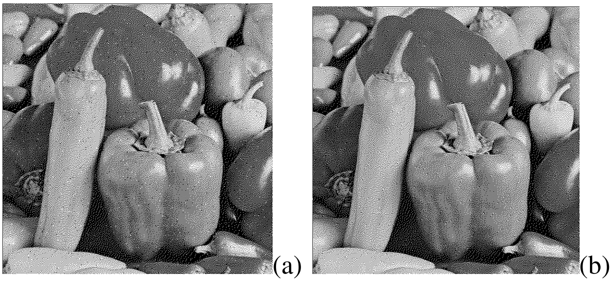


Fig. 6. The 4096 bits marked error diffusion Pepper

3.2 Tempered Region Recovery Capability

To evaluate tampered region recovery capability, the watermarked are altered intentionally. If the tempered region is within a rectangle of size less than sixteenth of the image, the proposed algorithm detects correctly the tampered region. The tempered pixels other than the pixels that are not used in encoding can be completely recovered. Fig.7 shows the tampered image, the authentication watermark and the watermark extracted from the tampered image. The tampered image is a 1/16 cropped marked Lena. The authentication watermark is of size 256x256. The cropped region is of size 128x128. At most one pixel in the cropped region will be included in each 15-bits code word. If the altered pixels are included in the code words, they can be recovered by the error correcting scheme of Hamming coding.

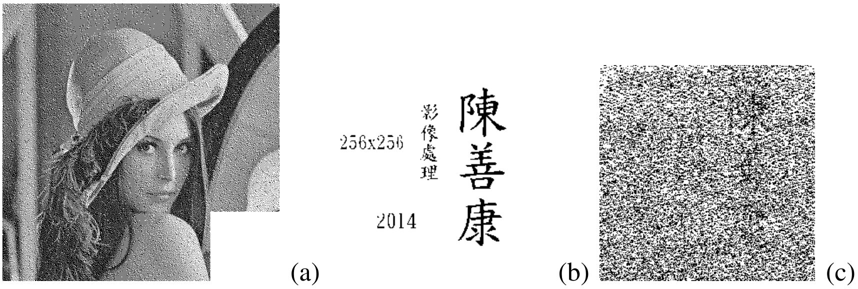


Fig. 7. (a) Tampered image (b) Authentication watermark (c) Watermark extracted from tampered image

4 Conclusion

In this paper, we propose a new data hiding scheme suitable for halftone images. We use the Space-filling curve to evenly split the original image into dispersed blocks. Data are hidden into each block using Hamming coding by exchanging pair of pixels selected through forward search. The exchanged pixels are neighboring pixels along Space-filling curve. Experimental results show that the proposed scheme dramatically decrease the chance of individual pixel toggling and thus greatly improve the visual quality of the marked halftone image. The error correcting ability of Hamming code is applied to locate and recover the tampered region.

References

1. Ulichney, R.: Digital Halftoning. The MIT Press, Cambridge (1987)
2. Wu, M., Liu, B.: Data Hiding in Binary Image for Authentication and Annotation. IEEE Trans. on Multimedia 6(4), 528–538 (2004)
3. Kim, H.Y.: A New Public-Key Authentication Watermarking for Binary Document Images Resistant to Parity Attacks. In: Proc. IEEE Int. Conf. on Image Processing, vol. 2, pp. 1074–1077 (2005)
4. Lien, B.K., Chang, C.Y.: Watermarking of Ordered Dither Halftone Images by Bit Interleaving. In: Proc. IEEE the 8th International Conference on Computer and Information Technology, vol. 1, pp. 319–322 (2007)
5. Baharav, Z., Shaked, D.: Watermarking of Dither Halftone Images. Hewlett-Packard Labs Tech. Rep., HPL-98-32 (1998)
6. Hel-Or, H.Z.: Watermarking and copyright labeling of printed images. J. Elect. Imaging 10(3), 794–803 (2001)
7. Pei, S.C., Guo, J.M., Lee, H.: Novel Robust Watermarking Technique in Dithering Halftone Images. IEEE Signal Processing 12(4), 333–336 (2005)
8. Guo, J.M., Pei, S.C., Lee, H.: Paired Subimage Matching Watermarking Method on Ordered Dither Images and Its High-Quality Progressive Coding. IEEE Trans. Multimedia 10(1), 16–30 (2008)
9. Fu, M.S., Au, O.C.: Halftone image data hiding with intensity selection and connection selection. Sig. Proc. Image Comm., 909–930 (2001)
10. Yu, F.-X., Luo, H., Chu, S.-C.: Lossless data hiding for halftone images. In: Pan, J.-S., Huang, H.-C., Jain, L.C. (eds.) Information Hiding and Applications. SCI, vol. 227, pp. 181–203. Springer, Heidelberg (2009)
11. Lu, Z.-M., Luo, H., Pan, J.-S.: Reversible watermarking for error diffused halftone image using statistical features. In: Shi, Y.Q., Jeon, B. (eds.) IWDW 2006. LNCS, vol. 4283, pp. 71–81. Springer, Heidelberg (2006)
12. Lien, B.K., Lin, Y.: High-capacity reversible data hiding by maximum-span pairing. Multimedia Tools and Applications 52, 499–511 (2011)
13. Pan, J.S., Luo, H., Lu, Z.M.: A lossless watermarking scheme for halftone image authentication. International Journal of Computer Science and Network Security 6(2b), 147–151 (2006)
14. Liao, P.-S., Pan, J.-S., Chen, Y.-H., Liao, B.-Y.: A lossless watermarking technique for halftone images. In: Khosla, R., Howlett, R.J., Jain, L.C. (eds.) KES 2005. LNCS (LNAI), vol. 3682, pp. 593–599. Springer, Heidelberg (2005)

15. Zhang, W., Wang, S., Zhang, X.: Improving embedding efficiency of covering codes for applications in steganography. IEEE Communication Letters 11(8), 680–682 (2007)
16. Bierbrauer, J., Fridrich, J.: Constructing good covering codes for applications in steganography (2008), `http://www.ws.binghamton.edu/fridrich/`
17. Kim, C., Shin, D., Shin, D.: Data Hiding in a Halftone Image Using Hamming Code (15, 11). In: Nguyen, N.T., Kim, C.-G., Janiak, A. (eds.) ACIIDS 2011, Part II. LNCS (LNAI), vol. 6592, pp. 372–381. Springer, Heidelberg (2011)
18. Fu, M.S., Au, O.C.: Data Hiding Watermarking for Halftone Images. IEEE Trans. Image Processing, 477–484 (2002)
19. Ma, Z., Li, F., Zhang, X.: Data Hiding in Halftone Images Based on Hamming Code and Slave Pixels. Journal of Shanghai University (Nature Science) 19(2), 111–115 (2013)
20. Moon, B., Jagadish, H.V., Faloutsos, C., Saltz, J.H.: Analysis of the Clustering Properties of the Hilbert Space-Filling Curve. IEEE Trans. Knowledge and Data Engineering 13(1), 124–141 (2001)

16-Bit DICOM Medical Images Lossless Hiding Scheme Based on Edge Sensing Prediction Mechanism

Tzu-Chuen Lu[1,*], Chun-Ya Tseng[1], Chun-Chih Huang[1], and Kuang-Mao Deng[2]

[1] Department of Information Management,
Chaoyang University of Technology, Taichung, 41349,Taiwan, R.O.C.
{tclu,s10033901}@cyut.edu.tw, opo1211@gmail.com
[2] Language Center, Chaoyang University of Technology, Taichung, 41349, Taiwan, R.O.C.
kmdeng@cyut.edu.tw

Abstract. Medical imaging is an important part of patient records. The pixel of a 16-depth DICOM image is totally different from the 8-bit depth nature image and is seldom the same as the other pixels in the nearby area. In this paper, we propose a reversible hiding method that expands Feng and Fan's prediction technique and adapts the scheme to match the characteristics of medical image. In the previous work, we determine what prediction method should be applied based on standard deviation thresholds to obtain more accurate prediction results. Finally, our approach includes embedding hidden information based on the histogram-shifting technique. The experimental results demonstrate that our approach achieves high-quality results.

Keywords: Medical imaging, Reversible data hiding, Standard deviation, Histogram shifting technique.

1 Introduction

Medical images are important data. Because Digital Imaging and Communications in Medicine (DICOM) constantly evolves, medical information can now be easily stored and sent via the Internet. Therefore, several data hiding techniques using digital medical images have been developed [1].

Data hiding techniques will induce some permanent destruction of the host image after embedding the hidden information. Therefore, several articles present research on reversible data hiding (RDH) to restore the hidden image to the original image without any loss whatsoever [5]. One approach is to use the differences between image pixels and the expansion prediction value to embed information [6]. One approach is to use the pixel distribution or the difference between prediction values and original pixels to analyze a histogram and embedded information in peak values [7]. In this paper, we use different prediction methods for pixels blocks in which the method used depends on where the standard deviation calculated for a block falls between two different thresholds. To reduce the artificial influence of threshold setting, this research utilizes an adaptive method to set the thresholds.

[*] Corresponding author.

H. Sun et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 329, DOI: 10.1007/978-3-319-12286-1_19

2 Related Works

Yang and Tsai proposed partitioning pixels into two sets, such as one resembling a checkerboard pattern, with hidden information determined based on a histogram shift technique [7]. Below are the equations defining the two pixel sets:

$$\begin{cases} \alpha = x_{(i,j)}, & if \quad (i\%2) = (j\%2), \\ \beta = x_{(i,j)}, & if \quad (i\%2) \neq (j\%2), \end{cases} \tag{1}$$

where $x_{(i,j)}$ is the original set of pixels, and i and j represent the pixel position in a two-dimensional image. Let α = $\{x_{(0,0)}, x_{(0,2)}, \ldots, x_{(m\text{-}1,n\text{-}1)}\}$ and β = $\{x_{(0,1)}, x_{(0,3)}, \ldots, x_{(m\text{-}1,n\text{-}2)}\}$, where $\alpha \cap \beta = \phi$.

Lukac et al. proposed a prediction approach through neighboring pixels to calculate edge-sensing weight coefficients [4]. They utilized neighboring pixels to calculate weight coefficients through the following formula:

$$p_{(i,j)} = \left\lfloor \sum_{k=1}^{4} w_k r_k \right\rfloor, \tag{2}$$

where $p_{(i,j)}$ is the prediction value, w_k is the weight of neighboring pixels, r_k is the k^{th} neighboring pixel, i and j are the pixel positions in the two-dimensional image, and k is the position set of neighboring pixels where $k = \{(i,j-1), (i-1,j), (i,j+1), (i+1,j)\}$. Through equation (2), we calculate weight w of position set k so that the prediction value $p_{(i,\ j)}$ is the sum of $w_k r_k$. Edge-sensing coefficients are utilized to calculate weight w_k through the formula below:

$$u_l = \frac{1}{1 + \sum_{k=1}^{4} |r_l - r_k|}, \tag{3}$$

where $l = \{(i,j\text{-}1), (i\text{-}1,j), (i,j+1), (i+1,j)\}$. Using equation (3) to determine the degree of difference among neighboring pixels, if r_l is greater than r_k, u_l will be close to zero. Otherwise, if r_l is smaller than r_k, then u_l will be close to one. Finally, we obtain w_k of the 4-neighbor pixels by calculating the normalized value of u_k with the following formula:

$$w_l = \frac{u_l}{\sum_{k=1}^{4} u_k}. \tag{4}$$

In 2012, Feng and Fan applied Lukac's method to hide data [3]. They assume 4-neighbor pixels $r_{(i,j-1)}$, $r_{(i-1,j)}$, $r_{(i,j+1)}$, and $r_{(i+1,j)}$ have different contribution to prediction value $p_{(i,j)}$. They use some additional pixels to compute the contribution of each neighboring pixel in the prediction process. First, the prediction value for the 4-neighbor pixels of $r_{(i,j)}$ is calculated as described above to generate $p'_{(i,j-1)}$, $p'_{(i-1,j)}$, $p'_{(i,j+1)}$, and $p'_{(i+1,j)}$. Next, the u_k and w_k via equations (3) and (4) are calculated for the 4-neighbor pixels, but the set k of neighbor pixels is replaced by $f_k = \{(i-1,j-1), (i-1,j+1), (i+1,j-1), (i+1, j+1)\}$. Finally, the effect size of predicted value p_k is adjusted with the following formula:

$$p'_k = \lfloor x_k - \eta_k \times (p_k - x_k) \rfloor, \tag{5}$$

where η_k is a parameter used to modify the effect size of the predicted value. The authors suggested the value of η_k should be $\eta_k = 0.1$.

3 Proposed Method

Our research applies an improvement to the Feng and Fan method to improve the accuracy of predictions. In our work, calculating standard deviations for 3×3 blocks is our approach to prediction. Further, we embed information based on histogram shifting. Fig. 1 shows the proposed process.

3.1 Preprocess Procedure

Initially, mimicking a checkerboard, all pixels are partitioned into a black squares set α and a white squares set β that are disjoint to one another.

Next, the image is divided into 3×3 blocks and the priority for each block is computed. In this paper, we use the standard deviation to order the priority according to the experimental results of Al-Qershi and Khoo [2].

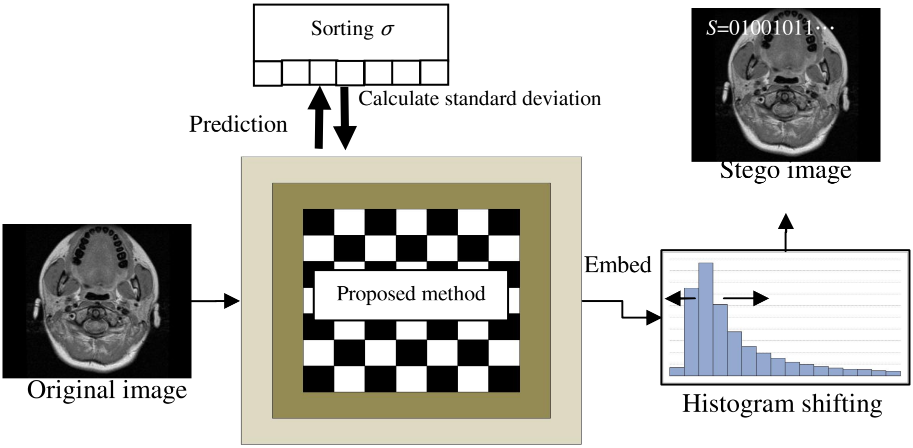


Fig. 1. The proposed process

Hence, $x_{(i,j)}$ of 4-neighbor pixels is used to calculate standard deviation, the formula is as follows:

$$\overline{r} = \frac{\sum_{c=1}^{4} r_c}{4}, \tag{6}$$

$$\sigma = \sqrt{\frac{1}{4}\sum_{c=1}^{4}\left(r_c - \overline{r}\right)^2}, \tag{7}$$

where $\overline{r}$ is the mean of neighbor pixels, c is the set denoting the position of neighbor pixels $c = \{(i-1, j), (i, j-1), (i, j+1), (i+1,j)\}$, and σ is the standard deviation. We sort the standard deviation values in ascending order and embed information into the lower σ values, because if σ is lower, the block is smoother and more suitable for embedding hidden information. Further, σ will have a lower impact on mean square error (MSE). If it is very high, the block is not considered suitable for embedding information.

3.2 Prediction Procedure

Our prediction method is determined by standard deviation thresholds T_1 and T_2. If σ is lower than T_1, the block is denoted as smooth and Lukac's approach is used for prediction. If σ is between T_1 and T_2, the block is denoted as complex and the expanded Lukac's approach is used for prediction. Otherwise, the block is denoted as very complex and therefore not suitable for embedding information.

3.3 Embedding Procedure

Ni's approach utilizes pixels of the entire image to build the histogram table. In our work, we build the histogram table based on prediction error $e_{(i,j)}$ and embed information into the peak point. The embedding process is described as follows:

(1) Build a prediction error $e_{(i,j)}$ histogram table.
(2) Determine peak point P and zero point Z from the histogram table.
(3) Modify error values between P and Z based on the following conditions:
I. If $P < Z$, then error value $e_{(i,j)}$ is increased by 1 between $P + 1$ and Z.
II. If $P > Z$, then error value $e_{(i,j)}$ is decreased by 1 between $P - 1$ and Z.
(4) Embed information into P and generate the stego image.

3.4 Extraction Procedure

The extraction process is defined as follows:

(1) Partition pixels into two disjoint sets α and β.
(2) Obtain standard deviations and sort these in ascending order.

(3) Receive the two threshold values from the embedding procedure and build a histogram based on prediction errors $e_{(i,j)}$.

(4) Receive peak point P and zero point Z from the embedding procedure, and then perform the extraction process described below:

   I. When $P < Z$:
   - a. If $e_{(i,j)} = P + 1$, then $e_{(i,j)}-1$ and extraction information is 1.
   - b. If $e_{(i,j)} = P$, then $e_{(i,j)}$ is unchanged and extraction information is 0.
   - c. If $e_{(i,j)}$ is between $P + 1$ and Z, then recover error values $e_{(i,j)}-1$

   II. When $P > Z$:
   - a. If $e_{(i,j)} = P-1$, then $e_{(i,j)}+1$ and extraction information is 1.
   - b. If $e_{(i,j)} = P$, then $e_{(i,j)}$ is unchanged and extraction information is 0.
   - c. If $e_{(i,j)}$ is between $P-1$ and Z, then recover error values $e_{(i,j)}+1$.

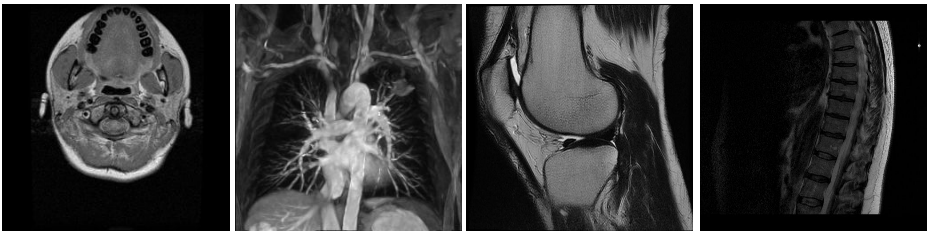
(5) Original pixels can be reconstructed from $x_{(i,j)} = p_{(i,j)} - e_{(i,j)}$.

4 Experimental Results

Matlab 7.10.0 (R2010a) is utilized to implement our experiments. Further, we use 16-bit DICOM medical images from the aycan OsiriX^{PRO} (http://www.aycan.de/lp/sample-dicom-images.html). Example images are shown in Fig. 2. The comparisons utilize PSNR measures to evaluate performance. Below is the PSNR formula:

$$PSNR = 10 \times \log_{10}\left[\frac{(ImageDepth)^2}{\frac{1}{mn}\times\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\left(x_{(i,j)} - x'_{(i,j)}\right)^2}\right] (\mathbf{dB}), \tag{8}$$

where *ImageDepth* is image depth 2^{bit}-1, $x_{(i,j)}$ is the set of original pixels, and $x'_{(i,j)}$ is the set of modified pixels. *ImageDepth* changes with image depth. For example, if image depth is 16 bits, then *ImageDepth* is $2^{16}-1=65{,}535$.

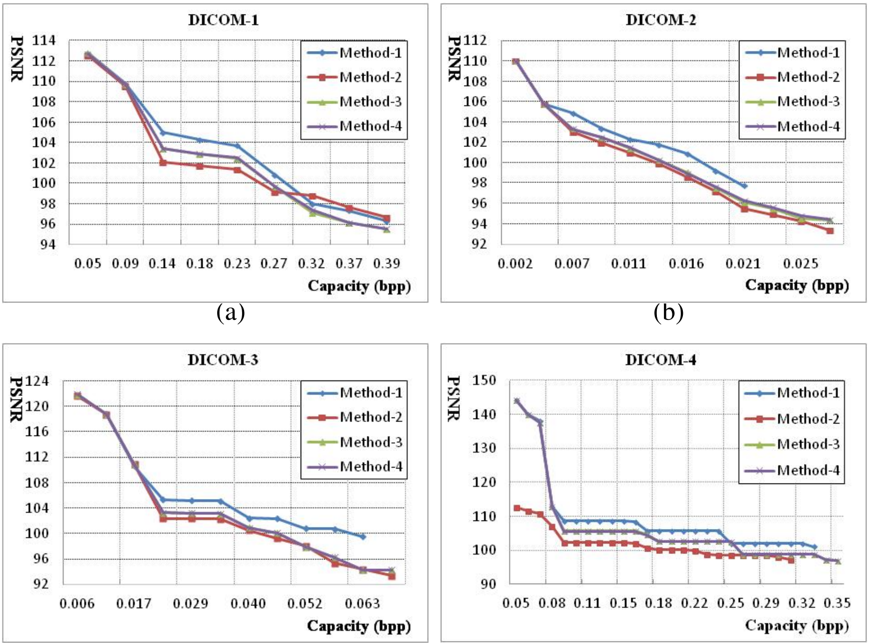


DICOM-1 (256x256) DICOM-2 (256x256) DICOM-3 (512x512) DICOM-4 (512x512)

Fig. 2. Test images used in our experiments: medical images

4.1 Adaptive Determination of Thresholds

We analyze standard deviations for each image. In our experimental results, the peak point of the standard deviation values is between 1 and 5 for most images. If the standard deviation is lower, then the block is smooth and more suitable for prediction via Lukac's traditional approach. Therefore, we determine T_1 from the peak point value. Next, threshold value T_2 is set according to the total hiding capacity. Assuming an image size of 512 × 512, the size of the hidden capacity is 260,100, excluding border pixels. For example, if user requirements call for a total hidden capacity of 80%, then T_2 should be selected to include 80% or more of the total hidden capacity.



(a) (b)

(c) (d)

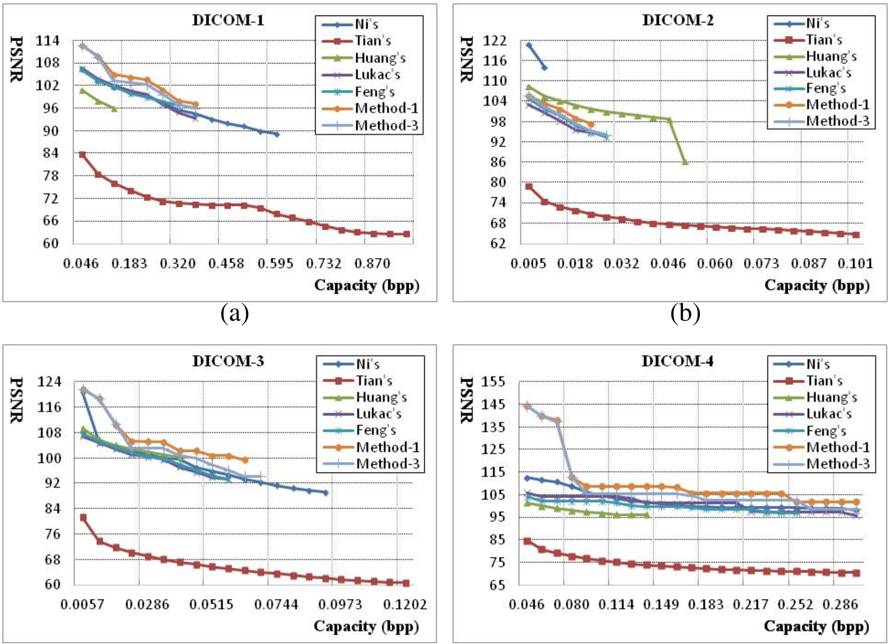
Fig. 3. Four iterations comparing hidden capacity and PSNR for different threshold-determination methods

Our proposed method can embed multiple pieces of information in the same image. Different thresholds result in different hidden results. Therefore, we use four methods to determine T_1 and T_2. The four methods are described below:

(1) Method-1: Entire image to generate thresholds, and utilize all the embed turns.
(2) Method-2: Entire image to generate T_1, and set threshold value T_2=100%.
(3) Method-3: Recalculate thresholds for each embed turn.
(4) Method-4: Calculate thresholds in the first time, and then use calculated thresholds for all embed turns.

4.2 Hidden Capacity and PSNR

This experiment compares our proposed method with those of other authors. Further, we select Method-1 and Method-3 (from Section 4.1 above) to determine thresholds. is used to compare hidden capacity and PSNR after four embedding iterations of the given DICOM images.



(a) (b)

(c) (d)

Fig. 4. Comparison of hidden capacity and PSNR

Overall, the DICOM images have larger background regions (excluding DICOM-2) and the peak point is zero, requiring little modifications based on histogram shift because 16-bit image characteristics make less difference distances between peak point and zero point. Tian's method has the highest embedding capacity. Because the 16-bit image only has the problem of underflow, too many conditions are reached to meet embedding and cause significant distortion. Huang's method is not as efficient as other methods, but his method has higher embed capacity based on PSNR = 86.32 dB when the image has more light block, such as DICOM-2 image. Lukac's method has more hidden capacity based on similar PSNR values than the expanded Lukac's method but is less efficient when the image has more complex blocks, such as DICOM-2 image. Our proposed method has more hidden capacity based on PSNR values of 95 dB or more.

In our experiments, we discovered that the image will be determined PSNR≧90 dB when the image has more gray area, such as scan images of the brain, spine, cells, etc. When PSNR≧80 dB, the image has imperceptibly been modified when the image has more lighter areas (such as with X-rays or visceral scans).

5 Conclusion

In this paper, we presented an improved Feng and Fan's approach. First, we used two thresholds that are generated through standard deviation analysis for each block and the statistical distribution of the standard deviation. Next, through the peak point of standard deviation, T_1 is determined and T_2 is inferred to limit the maximum embed capacity based on the ascending sorted order of the standard deviation. Experimental results indicate that this method can effectively improve PSNR and provide reasonable hidden capacity for 16-bits DICOM images.

References

[1] Al-Qershi, O.M., Khoo, B.E.: High Capacity Data Hiding Schemes for Medical Images Based on Difference Expansion. Journal of Systems and Software 84, 105–112 (2011)

[2] Al-Qershi, O.M., Khoo, B.E.: Two-Dimensional Difference Expansion(2D-De) Scheme with a Characteristics-based Threshold. Signal Processing 93, 154–162 (2013)

[3] Feng, G., Fan, L.: Reversible Data Hiding of High Payload Using Local Edge Sensing Prediction. Journal of Systems and Software 85, 392–399 (2012)

[4] Lukac, R., Martin, K., Plataniotis, K.N.: Digital Camera Zooming Based on Unified CFA Image Processing Steps. IEEE Transactions on Consumer Electronics 50, 15–24 (2004)

[5] Wen, J., Lei, J., Wan, Y.: Reversible Data Hiding Through Adaptive Prediction and Prediction Error Histogram Modification. International Journal of Fuzzy Systems 14(2), 244–256 (2012)

[6] Yang, W.J., Chung, K.L., Liao, H.Y., Yu, W.K.: Efficient Reversible Data Hiding Algorithm Based on Gradient-based Edge Direction Prediction. The Journal of Systems and Software 86, 567–580 (2013)

[7] Yang, C.H., Tsai, M.H.: Improving Histogram-Based Reversible Data Hiding by Interleaving Predictions. IET Image Processing 4(4), 223–234 (2010)

A Quality Improving Scheme for VQ Decompressed Image Based on DCT

Yung-Chen Chou, Shu-Huan Chen, and Min-Rui Hou

Department Computer Science and Information Engineering, Asia University,
No. 500, Lioufeng Rd., Wufeng, Taichung 41354, Taiwan, R.O.C.
{yungchen,gjcj06810622,g81lsui}@gmail.com
http://www.asia.edu.tw

Abstract. Data compression has been at an important stage, which not only needs to achieve higher compression ration but also needs to achieve the low distortion rate. High compression can make us be able to save the same data with smaller space; it can also be save the bandwidth of data transmission on networks. The proposed method is tried to further reduce the size of digital image and improve the visual quality of the decompressed image. The experimental results shows that the proposed method has better visual quality than VQ in case of AC codebook size is greater than 1024. On the other hand, the compression rate of VQ is 0.06 and the proposed method is 0.04 when AC codebook size set as 1024.

Keywords: Image compression, Vector Quantization, Discrete Cosine Transformation, Visual Quality.

1 Introduction

In the recent years, image compression becomes a popular research topic because the advanced development of information technique. Digital creation is easy to generate than before, thus how to significantly reduce the size of digital creation become a serious problem. Image compression is one of most useful strategy for saving the storage cost and network bandwidth. Image compression can briefly classified into two categories: Lossy compression and Lossless Compression [1,5,8,10,12,13,14,15].

The lossy compression means that the decompressed image is different from the original image. The benefit of lossy image is significantly reduce the size of digital image with the small distortion. VQ and JPEG are well-known lossy image compression algorithms. Contrary, lossless image compression more concentrate on the visual quality than saving the size of digital image. The decompressed digital image is fully the same with its original image. In some applications (e.g. medical treatment and military image), lossless property is fundamental requirement. But, in daily life image, tiny distortion on the image is acceptable because it is hard to distinguish the distortion on the image by using humans' eyes [9,11].

H. Sun et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 329, DOI: 10.1007/978-3-319-12286-1_20

Teskouras [2] and Nandi et al. [3] proposed fuzzy clustering methods to training codebook to improve quality of decompressed image. Further, Chen et al. propose post-processing technology after decompression of image [4] to improve the quality. Shen and Huang proposed an Adaptive Image Compression method, which utilize VQ encoding image and original to achieve improving the image quality [6]. Training coding book difference value method proposed by Shen and Lo is in contrast to traditional coding book training [7].

A wonderful lossy compression method is tried to significantly reduce the size of digital image and keep good visual quality of decompressed image. In this paper, we proposed a DCT based VQ compression method to further reduce the size of digital image also improve the visual quality of the decompressed image. The main idea of the proposed method is to transform the image into frequency domain and applying VQ compression strategy to generate the DC and AC index tables. Because DC values keeps the most important information of the digital image content, thus the DC difference between codeword and original DC values were concatenate with the index tables to form the final compression code.

The remaining sections are organization as follows. Some background knowledge were described in Section 2. After that, the key steps of the proposed image compression/decompression procedure were detailed in Section 3. Section 4 summarized the performance evaluation results. Finally, some conclusions are made in Section 5.

2 Related Works

In this section, we will introduce some background knowledge related to the proposed method such as Vector Quantization Coding, LBG algorithm, and Discrete Cosine Transform. Subsection 2.1, we introduce the main idea of Vector Quantization Coding. Then, a famous codebook training algorithm LBG is described in Subsection 2.2. Discrete Cosine Transformation is a very useful signal processing tool which can also be used in digital image processing. The main idea of DCT is detailed in Subsection 2.3.

2.1 VQ

Vector quantization (VQ) is a widely used concept for many applications that is presented by Y. Linde, A. Buzo and R. M. Gray in 1980 [16]. The main idea of VQ compression is to remember the most similar codeword from a prepared codebook for every image blocks. Thus, the index table is the compression code of image. First, the image I is divided into non-overlapping blocks sized $h \times w$. Then, for every block, find out the most similar codeword from codebook CB. Let the VQ codebook denoted as $CB = \{cw_i | i = 0, 1, \ldots, N-1\}$ and $cw_i = \{x_j | j = 0, 1, \ldots, h \times w - 1\}$. The most similar codeword means that there has smallest Euclidean distance between image block and the codeword. Where the Euclidean distance is defined as Eq. 1.

$$d(B_i, cw_j) = \sqrt{\sum_{k=0}^{h\times w-1} (b_k - x_k)^2} \tag{1}$$

Where b_k and x_k denoted as the k-th pixel in the block B_i and codeword cw_j, respectively. Fig. 1 illustrates the concept of the VQ compression. The VQ decompression is a reverse work which reconstruct the image by referring codeword indexes and use corresponding codeword to reconstruct the image.

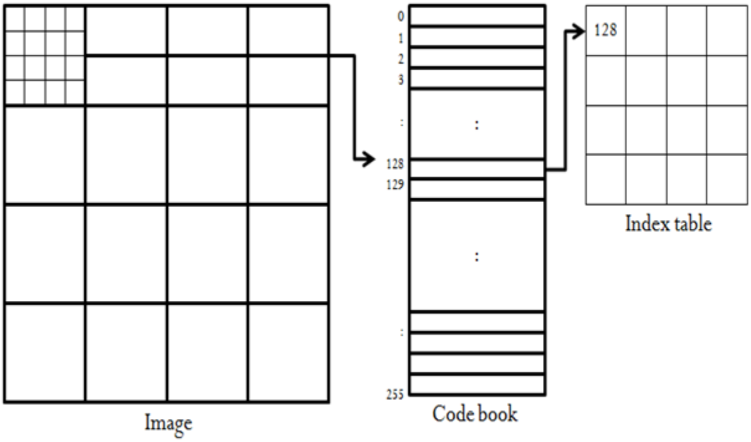


Fig. 1. VQ compression process

2.2 LBG Algorithm

As mentioned above, VQ codebook is a very important component for achieve the goal of image compression. The codebook training also an important research issue. LBG algorithm is a popular codebook training method which is proposed by Y. Linde, A. Buzo and R. M. Gray in 1980. First, select some images to be the training images. Then, divide training image into non-overlapping blocks to form the training pool. After that, randomly choice N blocks from training pool to form an initial codebook. Then, each block is classified into one of initial codeword group. After all of image blocks in the training pool have been classified into codeword group, to calculate the central vectors of codeword group to be the new generation codebook. The final codebook is generated when the new central vectors are mostly similar to previous generation.

2.3 Discrete Cosine Transformation

The DCT is a signal procession function, which can also be applied in two dimensional image processing. The 2D DCT transformation includes FDCT (Forward

Discrete Cosine Transformation) and IDCT (Inverse Discrete Cosine Transformation) which used to transform spatial domain to the frequency domain and from frequency domain to spatial domain. The 2D DCT transformation is performed on a block unit containing $h \times w$ pixels. The FDCT and IDCT functions are defined in Eqs. 2-3, respectively.

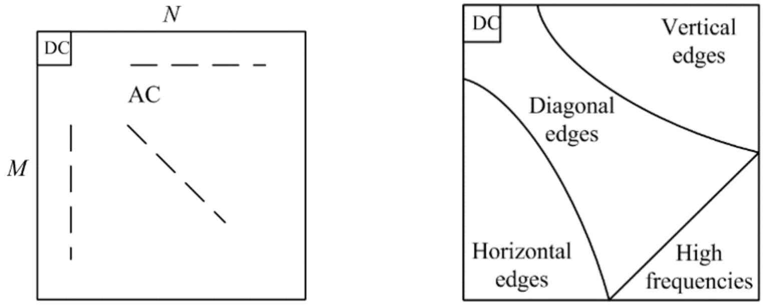
$$F(u,v) = \frac{1}{4}\alpha_u\alpha_v \sum_{m=0}^{h-1}\sum_{n=0}^{w-1} f(m,n) \times \cos\left[\frac{(2m+1)\pi u}{2h}\right]\cos\left[\frac{(2n+1)\pi v}{2w}\right] \quad (2)$$

$$f(m,n) = \frac{1}{4}\sum_{u=0}^{U-1}\sum_{v=0}^{V-1} \alpha_u\alpha_v F(u,v) \times \cos\left[\frac{(2m+1)\pi u}{2h}\right]\cos\left[\frac{(2n+1)j\pi v}{2w}\right] \quad (3)$$

$$\alpha_u = \begin{cases} \frac{1}{\sqrt{2}}, & u = 0 \\ 1, & 1 \leq u \leq h-1 \end{cases} \quad (4)$$

$$\alpha_v = \begin{cases} \frac{1}{\sqrt{2}}, & v = 0 \\ 1, & 1 \leq v \leq w-1 \end{cases} \quad (5)$$

Where $F(u,v)$ represents the coefficient value located at (u,v). The notation $f(m,n)$ represents a pixel value located at coordinate (m,n) in a block. Fig. 2) demonstrate the energy distribution of image content. The DC coefficient located at the left top corner. The remaining coefficients in the block also call the AC coefficients (i.e., referring to Fig. 2(a)). Where, DC is the most important value of the coefficient block that collecting the most information of image content (i.e., referring to Fig. 2(b)). The AC coefficient is related to the detail of image content.



(a) DC and AC coefficients distribution (b) Structural decomposition of coefficients

Fig. 2. The DCT coefficients distribution

3 Proposed Method

The main idea of the proposed method is using DCT codebooks to encode an image. First, the DCT codebook training is detailed in Subsection 3.1. Then, the proposed DCT based VQ compression and decompression were described in Subsections 3.2 and 3.3, respectively.

3.1 Codebook Training

As we know, the DCT coefficient composed by DC and AC coefficients, the DC value always significantly greater than AC coefficients. On the other hand, the DC value is the most important information of the block. Thus, the proposed method needs to train both DC codebook and AC codebook. We adopt LBG algorithm for both DC codebook and AC codebook training. First, five common used test images (i.e., Baboon, Boats, Goldhill, Lena, and Pepper) were selected to be the training images. Then, every training image is divided into non-overlapping blocks sized 8 × 8 pixels and transformed to DCT coefficient blocks.

For DC codebook training, all of DC values were collected to form training vectors (i.e., every vector composed by V_{DC} elements). For example, five training images sized 512 × 512, the block size is 8 × 8, and every training vector composed $V_{DC} = 8$ elements then the training pool must be a 2560 × 8 array. After that, the initial codebook is generated by randomly choose N_{DC} vectors from the training pool. According to LBG algorithm procedure, the DC codebook will be gained when the termination condition is satisfied.

For AC codebook training, the AC coefficients in a block is used to form one of vectors in the training pool. For example, five training images sized 512 × 512, the block size is 8 × 8, and every training vector composed $V_{AC} =$ 63 elements then the training pool must be a 20480 × 63 array. Again, the initial AC codebook is generated by randomly selected from the AC training pool. According to LBG algorithm procedure, the AC codebook can be gained when the termination condition is satisfied. Finally, the DC codebook $CB_{DC} = \{cw_i^{DC} | i = 0, 1, \ldots, N_{DC}\}$ and $CB_{AC} = \{cw_j^{AC} | j = 0, 1, \ldots, N_{AC}\}$.

3.2 Image Compression

For simplest, a gray scale image is represented as $I = \{p_{i,j} | i = 0, 1, \ldots, H - 1; j = 0, 1, \ldots, W - 1\}$ and $p_{i,j} \in \{0, 1, \ldots, 255\}$. Because the proposed image compression method is block wise thus the image can also be represented as $I = \{B_i | i = 1, 2, \ldots, N_B\}$ where B_i represent the i-th block sized $h \times w$ and N_B is the total number of blocks.

First, the image I is divided into non-overlapping blocks sized $h \times w$ pixels. For each block B_i is transformed to frequency domain by adapting DCT transformation function. After that, all of DC coefficients in the image is picked up to form the DC vectors. Then, for each DC vector, finding out the closest DC codeword from CBDC to form the DC index table IT_{DC}. On the other hand, the DC coefficients difference between original DC values and the corresponding codeword values were collected to form the extra data $Diff_{DC}$.

For the AC coefficients, every DCT block is encoded by figuring out the closest codeword from the AC codebook CB_{AC}. Here, the closest codeword selection is to find the smallest Euclidean distance between the AC coefficients in the block and the codeword in CB_{AC}. Finally, the compression code is generated by concatenate the DC index table, DC difference, and the AC index table. The key steps of the proposed image compression procedure are summarized as following procedure.

Image compression procedure

Input: image I and codebooks CB_{DC} and CB_{AC}
Output: the compression code I_c
Step 1: Divide I into non-overlapping blocks sized $h \times w$ pixels.
Step 2: Transform every block B_i into DCT coefficients by using FDCT function.
Step 3: Encode all of DC coefficients by using VQ encoding with DC codebook CB_{DC}.
Step 4: Calculate the difference between original DC values and DC codewords.
Step 5: Record the index of the selected DC codeword and the DC difference data.
Step 6: Encode all of AC coefficients by using VQ encoding with AC codebook CB_{AC}.
Step 7: Concatenate the DC indices, AC indices, and addition information to form the compression code.

3.3 Image Decompression

Assume that both of the encoder and decoder side have the same DC codebook and AC codebook. The image can be reconstructed by parsing the compression code and using the corresponding DC codewords and AC codewords. First, take $\lceil \log_2(N_{DC}) \rceil$ bits to form the DC index from the compression code. Then the DCs of blocks were filled back by the elements of the DC codeword. After all of the DC values have been filled backed, taking the $Diff_{DC}$ from the compression code and use it to adjust the DC values.

The remaining part of compression code is the AC codeword indexes. Every $\lceil \log_2(N_{AC}) \rceil$ is taken to be the index of AC codeword. The AC coefficients of the block B_i is filled back by using the corresponding codeword in CB_{AC}. After all of DCT coefficients have been reconstructed, the decompression image can be gained by applying IDCT (Inverse Discrete Cosine Transformation) function to transform the image from frequency domain to spatial domain. The key steps of the proposed image decompression procedure were summarized as follows:

Input: The compression code I_c
Output: The decompressed image I'
Step 1: Take $\lceil \log_2(N_{DC}) \rceil$ bits from I_c to form idx_{dc}

Step 2: Reconstruct image blocks by using the DC codeword corresponding to idx_{dc}.
Step 3: Take the $Diff_{DC}$ data from the compression code.
Step 4: Adjust DC values with $Diff_{DC}$ values.
Step 5: Take $\lceil \log_2(N_{AC}) \rceil$ bits from I_c to form idx_{ac}.
Step 6: Reconstruct B_i by using AC codeword corresponding to idx_{ac}.
Step 7: Repeat **Step 6** until all of blocks have been filled back the coefficients.
Step 8: Transform frequency domain image to spatial domain by using IDCT function.

4 Experiment Results

In order to evaluate the performance of the proposed method, we implement traditional vector quantization method and the proposed method using Octave software works on Ubuntu 14.04 operating system. Six common images sized 512 × 512 were used in our experimental namely Baboon, Boats, Goldhill, Lena, Pepper, and Zelda (i.e., referring to Fig. 3). The codebook training is applying LBG algorithm. Four different DC codebook sizes were generated that including 256, 512, 1024, and 2048. The vector of DC codeword is set to 8. We found that visual quality performance of the proposed method is no matter to the size of DC codebook. The reason is that the proposed method has encoding the DC differences, thus DC values can be closely reconstructed to the original value. Also, we generated five AC codebooks (i.e., 256, 512, 1024, 2048, and 4096) to test the visual quality performance of the proposed method.

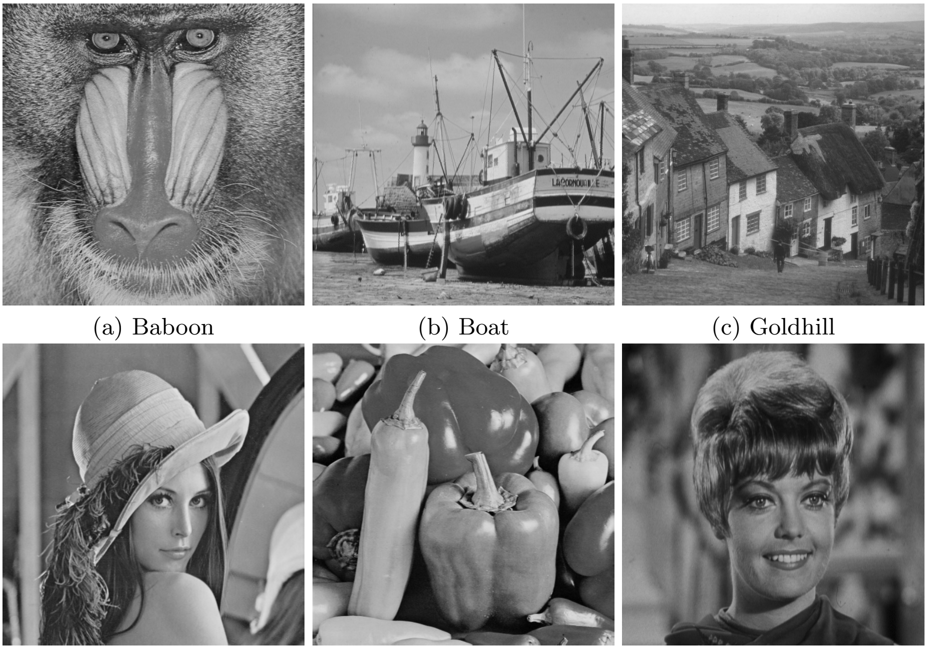
For evaluating the visual quality of the decompressed image, we adopt an objective measurement method the Peak-Signal-to-Noise-Ratio (PSNR) in our simulations. The PSNR is defined as Eqs. 6-7.

$$PSNR = 10 \times \log_{10} \frac{255^2}{MSE} dB \tag{6}$$

$$MSE = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} (p_{i,j} - p'_{i,j})^2. \tag{7}$$

Where H and W represent the height and width of the test image, respectively. $p_{i,j}$ and $p'_{i,j}$ denote to pixels in the original image and decompressed image, respectively. MSE represents the mean square error. A large PSNR value means that the decompressed image is most similar to the original one. In the other words, the visual quality of the decompressed image is good. Contrary, a small PSNR value indicates that the decompressed image has worse visual quality outcome.

Another important factor for evaluating the performance of image compression method is compression rate (CR). The compression rate is to calculate the total bits of the compression code and the original image. The formula is defined in Eq. 8. The notation $\|\cdot\|$ is the total bits of image or compression code.



(a) Baboon (b) Boat (c) Goldhill

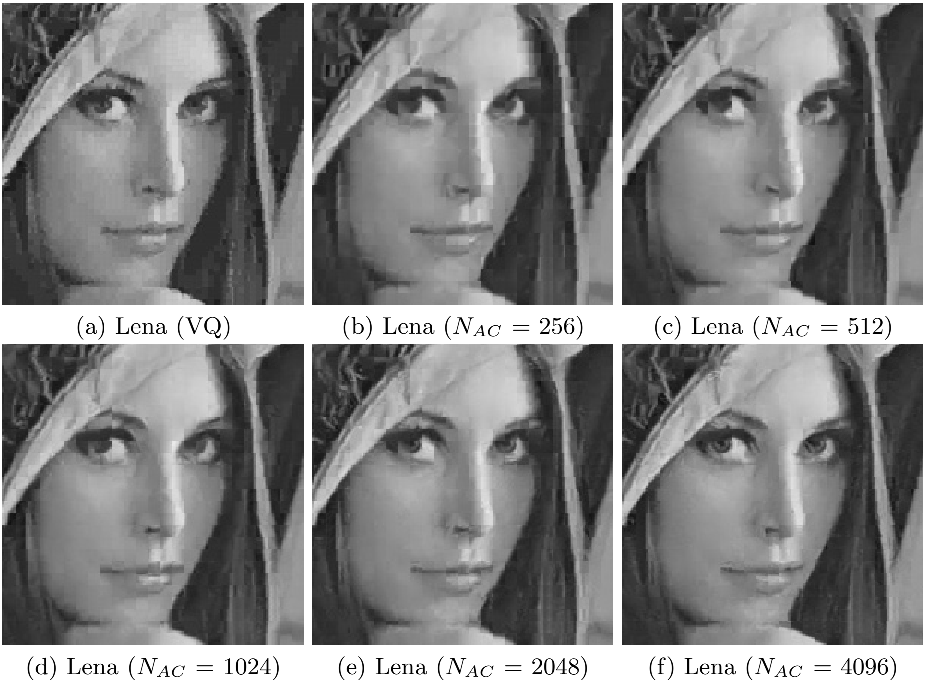
(d) Lena (e) Pepper (f) Zelda

Fig. 3. The test images sized 512 × 512

$$cr = \frac{\|I'\|}{\|I\|} \tag{8}$$

Fig. 4 demonstrate the visual quality comparison of VQ and the proposed method in different AC codebook size setting. As we can see, the visual quality of the proposed method is better than VQ in case of AC codebook size greater than 1024.

Table 1 summarized the comparison results of different AC codebook size setting. Experimental results show that the compression rate of the proposed method is better than VQ whether the size of AC codebook. In PSNR comparison, the visual quality of decompressed image which generated by the proposed method is worse than VQ when AC codebook size is smaller than 1024. We found that the proposed method has better performance than VQ in terms of visual quality when the AC codebook is larger than and equal to 1024. There has a flexible of the proposed method, if a user more consider the compression rate than visual quality, then using $N_{AC} = 1024$. Contrary, if a user more consider the visual quality of the decompressed image than compress rate, then using N_{AC} = 4096 AC codebook is a good choice. However, the proposed method provides better compression rate than VQ in any AC codebook size.


(a) Lena (VQ)

(b) Lena ($N_{AC} = 256$)

(c) Lena ($N_{AC} = 512$)

(d) Lena ($N_{AC} = 1024$)

(e) Lena ($N_{AC} = 2048$)

(f) Lena ($N_{AC} = 4096$)

Fig. 4. The visual quality comparison for different AC codebook setting

Table 1. The performance comparison in $N_{DC} = 256$ and different N_{AC} setting

	VQ		$N_{AC} = 256$		$N_{AC} = 512$		$N_{AC} = 1024$		$N_{AC} = 2048$		$N_{AC} = 4096$	
Images	PSNR	CR	PSNR	CR	PSNR	CR	PSNR	CR	PSNR	CR	PSNR	CR
Baboon	24.66	0.06	23.30	0.04	23.76	0.04	24.32	0.04	25.09	0.04	26.49	0.05
Boat	28.85	0.06	27.60	0.04	28.06	0.04	28.71	0.04	29.27	0.04	30.73	0.05
GoldHill	29.54	0.06	28.51	0.04	28.94	0.04	29.39	0.04	29.96	0.04	31.09	0.05
Lena	31.15	0.06	29.80	0.04	30.35	0.04	30.86	0.04	31.52	0.04	32.59	0.05
Pepper	30.94	0.06	29.88	0.04	30.47	0.04	31.05	0.04	30.21	0.04	32.83	0.05
Zelda	32.91	0.06	32.47	0.04	32.76	0.04	33.00	0.04	33.18	0.04	33.42	0.05

5 Conclusion

Vector quantization is a widely used concept in many applications such as introduction intrusion detection, and image compression. VQ image compression is a lossy compression that means the decompressed image is different from the original image. The proposed method tries to improve the visual quality of the decompressed image. The high computation cost is the weakness of the proposed method, which is not suitable in the mobile devices. However, the proposed method not only has better compression rate than VQ but also achieve the goal of visual quality improvement of decompressed image.

References

1. Chou, Y.C., Lo, Y.H., Shen, J.J.: A New Quality Improving Scheme for VQ Decompressed Images Based on DWT. Journal of Electronic Science and Technology 11(1), 51–57 (2013)
2. Tsekouras, G.E.: A Fuzzy Vector Quantization Approach to Image Compression. Applied Mathematics and Computation 167, 539–560 (2005)
3. Xu, W., Nandi, A.K., Zhang, J.: Novel Fuzzy Reinforced Learning Vector Quantisation Algorithm and Its Application in Image Compression. In: IEEE Proceedings Vision, Image, and Signal Processing, vol. 150, pp. 292–298 (2003)
4. Chen, S., He, Z., Luk, B.L.: A Generic Postprocessing Technique for Image Compression. IEEE Transactions on Circuits and Systems for Video Technology 11, 546–553 (2001)
5. Lancini, R., Tubaro, S.: Adaptive Vector Quantization for Picture Coding Using Neural Networks. IEEE Transactions on Communications 43, 534–544 (1995)
6. Shen, J.J., Huang, H.C.: An Adaptive Image Compression Method Based on Vector Quantization. In: Proceedings of the 1st International Conference Pervasive Computing Signal Processing and Applications, Harbin, China, pp. 377–381 (2010)
7. Shen, J.J., Lo, Y.H.: A New Approach of Image Compression Based on Difference Vector Quantization. In: Proceedings of the 7th International Conference Intelligent Information Hiding and Multimedia Signal Processing, Dalian, China, pp. 137–140 (2011)
8. Qian, S.E.: Hyperspectral Data Compression Using a FastVector Quantization Algorithm. IEEE Transactions on Geoscience and Remote Sensing 42(8), 1791–1798 (2004)
9. Liu, Y.C., Lee, G.H., Taur, J., Tao, C.W.: Index Compression for Vector Quantisation Using Modified Coding Tree Assignment Scheme. IET Image Processing 8(3), 173–182 (2014)
10. Chou, P.H., Meng, T.H.: Vertex Data Compression through Vector Quantization. IEEE Transactions on Visualization and Computer Graphics 8(4), 373–382 (2002)
11. Shen, G., Liou, M.L.: An Efficient Codebook Post-Processing Technique and a Window-Based Fast-Search Algorithm for Image Vector Quantization. IEEE Transactions on Circuits and Systems for Video Technology 10(6), 990–997 (2000)
12. Bagheri Zadeh, P., Buggy, T., Sheikh Akbari, A.: Statistical, DCT and Vector Quantisation-based Video Codec. IET Image Processing 2(3), 107–115 (2008)
13. Bayer, F.M., Cintra, R.J.: Image Compression Via a Fast DCT Approximation. IEEE Latin America Transactions 8(6), 708–713 (2010)
14. Ponomarenko, N.N., Egiazarian, K.O., Lukin, V.V., Astola, J.T.: High-Quality DCT-Based Image Compression Using Partition Schemes. IEEE Signal Processing Letters 14(2), 105–108 (2007)
15. Zhao, D., Gao, W., Chan, Y.K.: Morphological Representation of DCT Coefficients for Image Compression. IEEE Transactions on Circuits and Systems for Video Technology 12(9), 819–823 (2002)
16. Linde, Y., Buzo, A., Gary, R.M.: An Algorithm for Vector Quantization Design. IEEE Transactions on Commnunications 28(1), 84–95 (1980)

Research of Image Enhancement Algorithm Based on Multi-scale Retinex

Xinhua Yang[1] and Jianlong Yang[2]

[1] Hubei University of Technology, Wuhan, China
351732030@qq.com
[2] Nanchang Institute of Technology, Nanchang, China
ZG_tdcq@163.com

Abstract. Parking supplementary system has been such broadly used in some vehicle that we could drive and park more safely and conveniently. As light intensity has important effect on images from cameras which are mounted on vehicle's four different directions in order to obtain panoramic view. To eliminate the influence of light to the image and obtain consistent brightness of image, this paper proposed mix-algorithm combined histogram equalization (HE) with multi-scale retina (MSR). This method makes equalization processed image to MSR, and gets a enhanced result at last. This algorithm can decrease effectively noise and enhance the brightness of image. The performance of output image is shorter time, smaller mean square error and higher peak signal to noise ratio than MSR. The experimental results demonstrate that the proposed method can not only counterpoise illumination but also meet real-time requirements. So the proposed mix-algorithm can apply in lack of brightness and higher real-time such as parking supplementary system, traffic monitoring system and security systems etc.

Keywords: parking supplementary system, mix-algorithm, HE, MSR, real-time.

1 Introduction

Image enhancement, as front section of image process which is a very important step to better the performance of subsequent steps, has been broadly applied in many areas. There is no doubt that image enhancement has become such a significant branch that it attracts high research interest since the birth of image processing disciplines. Here we can briefly formulate the task of image enhancement as follows: Efficiently enhance details, information with 'interested' and remove harmful effect such as low illumination, strong illumination, noise, etc. It is expected that the image will be made 'comfortable' for computer or mankind after such operation is performed. Because the application of image enhancement algorithm is targeted, there is not universal enhancement method which is suitable to the all background of application. Up to now, numerous image enhancement techniques have been proposed by researchers to accomplish the task of improving image's quality. We can choose different enhancement algorithm that processed image is more suitable than original image for

H. Sun et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 329, DOI: 10.1007/978-3-319-12286-1_21

specific objective. Among them the simplest and most classical technique is histogram equalization which is a point processing method in the spatial domain [1, 2]. Its essence is to reduce the gray level of the image to increase the contrast. But it is not appropriate for color image because it is time-consuming for converting gray-scale image into color image. Homomorphic filtering based on image illumination/reflectance model is an approach to combine frequency filtering with gray-scale transformation [2]. This filter can be used to remove the effect of insufficient light and retain image detail simultaneously. But Fourier transform and inverse Fourier transform is time-consuming and its calculation needs more storage space, so this method is not suitable for practical image processing. Retinex theory is adopted by Land [3], which presents that sensations of color show a strong correlation with reflectance, even though the amount of visible light reaching the eye depends on the product of reflectance and illumination. The core idea of Retinex theory is that a image is composed of illumination image and reflectance image. We can achieve image enhancement by reducing the effect of reflectance image to illumination image.

There are lots of algorithms based on Retinex theory having been proposed. For example, Jobson et al. put forward the method of Center/Surround Retinex including single-scale Retinex(SSR)and MSR[4, 5]. Although this method increases the overall brightness of image, it also suffers the existence of halo artifacts and brings about noise. Kimmel et al. propose a variational model for the Retinex problem, as in [6]. Based on the variational model, this paper shows that the illumination estimation problem can be formulated as a quadratic programming optimization problem. Though this method can obtain better result, it is rarely used in practical application for consuming more time. In order to preserve the naturalness and enhance details, some natural enhancement methods based on Retinex theory are proposed to enhance details with the naturalness preserved simultaneously[7, 8]. However, these methods are not fit for uneven illumination. Recently, Wang et al. [9] propose a naturalness preserved enhancement algorithm for non-uniform illumination images. In this paper, since this algorithm does not take into consideration the relation of illumination in different scenes, it will introduce flickering for video applications. Li et al. propose a improved algorithm for remote sensing image based on MSR in 2014[10]. This improved algorithm presents a better result in suppressing image gradation, enhancing image space details and reducing color distortion. But this method applied for color image and grayscale image is not so good.

Therefore, we propose a mix-algorithm integrated HE and MSR. This method can decrease noise and enhance brightness of image through HE before the processing of MSR. For simple algorithm and fast efficiency, it is easily applied to the practical case, specifically to image processing.

The remainder of this paper is organized as follows. A exhaustive description of algorithm is presented in Section 2. In Section 3, some illustrative simulation examples are provided to test the algorithms, and we show the experimental results of using different algorithms and give image assessment criteria, which results verify the feasibility and real-time of the proposed method. Finally, Section 4 concludes the paper.

2 Algorithm Research

2.1 The Histogram Equalization

Image histogram describes the relative frequency of each gray level in image. Image can be adjusted to a predetermined shape based on the conversion of gray-scale. For example, the gray-scale of some images distribute in concentrated and narrower range, so it is weak contrast that leads to unclear details. We can adopt HE method to adjust the gray-scale of image, which can makes uniform distribution of gray-scale. And then visual effects of image can be improved, so it can achieve the purpose of image enhancement. Taking the visual characteristic of humankind consideration, this image tone is more coordinate if the histogram of image is uniform.

Assuming that the gray levels of original image(r) are normalized to between 0 and 1. i.e. r∊ [0,1], $r=0$ represent black and $r=1$ represent white, and gray level of image is [0, L-1]. $\rho_r(r)$ represents probability density function of gray distribution of the original image, the essence of HE is try to find a gray-scale transformation function T, let the gray value with transformed $s=T(r)$, where s is normalized to between 0 and 1, and then mapping relationship between r and s is established. But probability density function with processed image $\rho_s(s)$ is required to equal 1, that is, probability density with converted belongs $U(0,1)$. The transformation function of HE is shown in Fig.1.

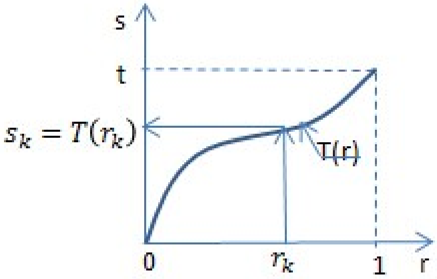


Fig. 1. Transformation function of HE

We can see from Fig.1 that the number of pixels is constant within the range of gray-scale transformation (Δr and Δs), so

$$\int_{r}^{r+\Delta r} \rho_r(r) \, dr = \int_{s}^{s+\Delta s} \rho_s(s) ds \tag{1}$$

. When $\Delta r \to 0$, $\Delta s \to 0$, we can get $\frac{ds}{dr} = \frac{\rho_r(r)}{\rho_s(s)}$.

Owing to s=T(r), $\rho_S(s)=1$, then $\frac{ds}{dr}=\frac{dT(r)}{dr}=\rho_r(r)$ .At last, the gray-scale transformation function of HE is shown as followed:

$$s=T(r)=\int_0^r \rho_r(r)dr \tag{2}$$

That is also cumulative distribution function of the original image gray level r. The calculation steps of HE is as followed for digital image:

Count the Histogram of the Original Image, $\rho_r(r_k)=\frac{n_k}{n}(0\le k\le L-1);$.

where r_k represents the normalized gray level of the input image, n_k is the number of pixels when image gray-scale level is r_k and n is the number of total image pixels.

Calculate the Cumulative Distribution Curve of Histogram, $s_k=T(r_k)=\sum_{j=0}^{k}\rho_r\left(r_j\right)=\sum_{j=0}^{k}\frac{n_j}{n}$.

Use (2) to Convert Image Gray-Scale. According to calculated cumulative distribution function, we establish corresponding relationship between the gray level of input image and output image. That is to say, cumulative distribution function s is repositioned. The enhanced results of HE contains two aspects. Enhanced image histogram tends flat and its gray-scale level is merged. The interval of several gray-scale level with original image contained more pixels is widened, but compress the interval of several gray-scale level with little pixels. So the amount of information received from vision systems is greatly enhanced.

2.2 The Multi-scale Retinex

The multi-scale Retinex(MSR)[5] is proposed by Zia-ur Rahman et al. based on single-scale Retinex (SSR) described in detail[3]. The advantage that the MSR has over the SSR is in the combination of scales which provide both tonal rendition and dynamic range compression simultaneously. MSR synthesizes several SSR, which mathematical expressions can written as

$$R_i(x,y)=\sum_{n-1}^{N}W_n\left\{\log\left[s_i(x,y)\right]-\log\left[F_n(x,y)*S_i(x,y)\right]\right\} \tag{3}$$

where the subscripts i ϵ R, G, B represent the three color bands, N is the number of scales being used, “*”denotes the convolution operation. w_n are the weighting factors for the scales, $\sum_{n=1}^{N}W_n=1$, MSR change into SSR when N=1, and $w_1=1$. S(x, y) represents original image, R(x, y) represents enhanced image. The $F_n(x,y)$ are the surround functions given by

$$F_n(x,y) = K_n e^{-(x^2+y^2)/\sigma^2 n} \tag{4}$$

where k_n is normalization factor and it is selected so that $\iint F(x,y)=1$, $\sigma_n{}^2$ are the standard deviations of the Gaussian function that determine the size of scale. In fact, we can select different scale to our needs. For example, the larger scale represents more color constancy, and smaller scales provide more dynamic range compression. In order to trade-off between dynamic range compression and color rendition, $\sigma_n = 80$ is proved to appropriate[4].

2.3 The Hybrid Algorithm

The MSR produce most of the detail in the black areas but at the expense of strengthening noise in these regions. This noise lead to the poor signal to noise ratio(SNR). For it, we produce improved MSR algorithm that we make the output of HE as input of MSR. It can increase the contrast of the image through HE method, which make subsequent processing better and decrease effectively noise. The following is the flow char t of the hybrid algorithm.

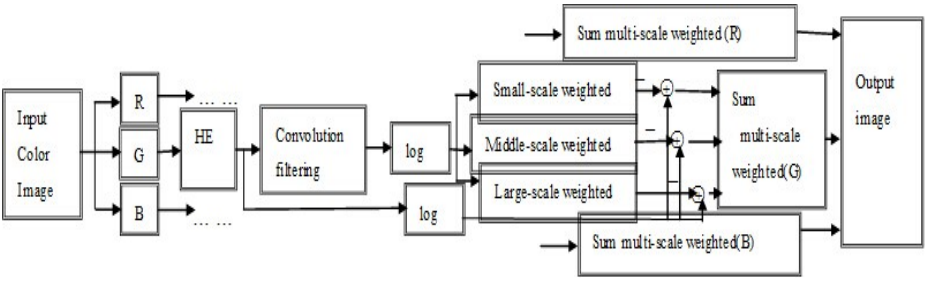
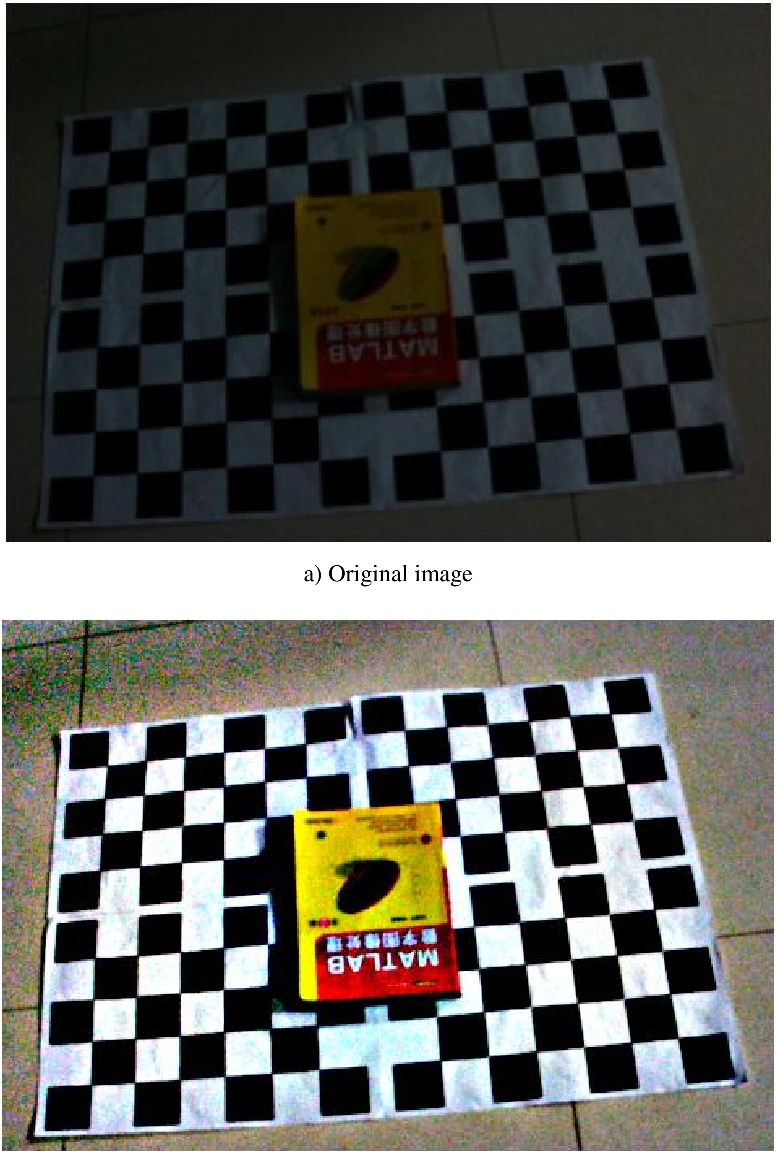


Fig. 2. Low chart of the hybrid algorithm

3 Experiment and Discussion

3.1 Experiment

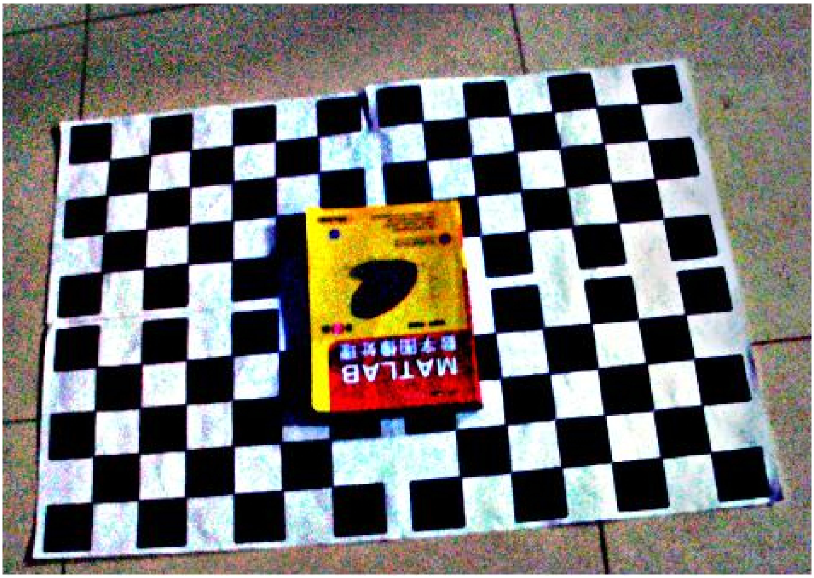
The proposed algorithm has been tested on our image which is from dark background, and the size of this image is 476*335. This image is a combination of practical application, for example, we drive at night or in the tunnel. These images often are low contrast and unclear detail, so we propose a hybrid algorithm combined HE and MSR. In order to verify the performance of the algorithm, it has been tested on our image in PC(Intel Core2, 2.66GHz, 2.00GB, Microsoft Windows XP, MATLAB 2010b). Fig.3 shows the result of experiment.



a) Original image

b) MSR

Fig. 3. Comparison of the different algorithms



c) Hybrid algorithm

Fig. 3. (*continued*)

3.2 Assessment Criteria

The assessment methods of image quality can be divided into subjective evaluation and objective evaluation. Subjective assessment mainly depends on human visual system, different levels, including 'excellent', 'good', 'middle' and 'terrible' is assessed by people. Objective assessment is often used to explain some important characteristics of the image. At the same time, it is the most widely used and accurate assessment methods. Common evaluation indicators include mean absolute error, mean square error (MSE), normalized mean square error, signal to noise ratio and peak signal to noise ratio (PSNR). Until now, there is also no uniform evaluation criteria to different enhancement algorithm. This paper selects MSE and PSNR to assess proposed algorithm.

The Smaller Is the Value of MSE, the Better Is Image Quality. And the calculation expression is as followed:

$$\mathrm{MSE} = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N}\left[f(i,j) - \hat{f}(i,j)\right]^2}{M * N} \tag{5}$$

where M and N represent the length and width of the image, respectively. And $f(i,j)$ and $\hat{f}(i,j)$ represent gray value of original image and output image when is (i, j), respectively.

The PSNR Describes the Ratio Useful Information to Overall Image. If PSNR is more larger, the quality of image is better. Its formula is as followed:

$$\mathrm{PSNR} = 10 * \log_{10}\left[\frac{255^2 * M * N}{\sum_{i=1}^{M}\sum_{j=1}^{N}\left[f(i,j) - \hat{f}(i,j)\right]^2}\right]. \tag{6}$$

Because real-time is a very important factor in practical application, so time, MSE and PSNR are listed in the following table.

Table 1. Quantitative evaluation table of image enhancement effects

•	• Time[s]	• MSE	• PSNR
– MSR	– 0.7350	– 3.1023	– 11.433
• Hybrid algorithm	• 0.7474	• 3.0598	• 11.554

We can see from the Table 1 that hybrid algorithm is better performance than MSR because the later includes more useful information and possess better image quality. Because this hybrid algorithm can balance noise between brightness, at the same time, and the result is suitable to the human visual system. It is applied in parking supplementary system which can presents more better result than MSR.

4 Conclusion

This paper proposes a mix algorithm combined HE and MSR, which not only strengthen the details of the image but also enhance the contrast. Experimental results show that the images enhanced by the proposed algorithm are real-time, more higher PSNR value and less MSE value. However, since our enhancement algorithm does not take illumination in different scenes into consideration, it may introduce different effects. At the same time, owing to bigger date information in image processing, there need faster process. So these will be our future research works.

Acknowledgment. First of all, I would like to extend my sincere gratitude to my supervisor, Liu Chun, for her instructive advice and useful suggestions on my paper. High tribute shall be paid to Yang Shi, whose profound knowledge of English triggers my love for this beautiful language and whose earnest attitude tells me how to learn English. Special thanks should go to my partner who have put considerable time and effort into this paper. Besides, I am indebted to my parents and girlfriend for their continuous supports and encouragement. Finally, this work was supported by the Graduate Innovative Fund Program of Nanchang Institute of Technology(item number: 2014yCXJJ-A2-001). We are very grateful to school for our supports and attentions.

References

1. Gonzalez, R.C., Woods, R.E.: Digital Image Processing, 2nd edn. Electronic Industry Press, Beijing (2011)
2. Jie, Y., Chaobing, H.: Digital Image Processing and MATLAB Achievement, pp. 74–98. Electronic Industry Press, Beijing (2013)
3. Land, E.H., McCann, J.J.: Lightness and retinex theory. J. Opt. Soc. Amer. 61(1), 1–11 (1971)
4. Jobson, D.J., Rahman, Z., Woodell, G.A.: Properties and performance of a center/surround retinex. J. IEEE Trans. Image Process. 6(3), 451–462 (1996)
5. Rahman, Z., Jobson, D.J., Woodell, G.A.: Multi-scale retinex for color image enhancement. In: Proc. Int. Conf. Image Process., pp. 1003–1006 (September 1996)
6. Kimmel, R., Elad, M., Shaked, D., Keshet, R., Sobel, I.: A variational framework for retinex. J. International Journal of Computer Vision 52(1), 7–23 (2003)
7. Chen, S., Beghdadi, A.: Natural enhancement of color image. EURASIP J. Image Video Process. 2010, 175203-1–175203-19 (2010)
8. Li, B., Wang, S., Geng, Y.: Image enhancement based on Retinex and lightness decomposition. In: Proc. IEEE Int. Conf. Image Process., pp. 3417–3420 (September 2011)
9. Shuhang, W., Jin, Z., Haimiao, H., Bo, L.: Naturalness preserved enhancement algorithm for non-uniform illumination images. J. IEEE Trans. Image Processing 22(9) (September 2013)
10. Jianchun, L., Limei, Z., Jun, L.: Algorithm for Remote Sensing Image Enhancement Based on Multi-scale Retinex Theory. Journal of Xi'an Technological Univercity 34(1) (January 2014)

The HRV Analysis on e-Book Reading

Chien-Sheng Huang*, Ching-Huang Lin, Chien-Yue Chen,
Shao-Ciang Gan, and Guan-Syuan Hong

National Yunlin University of Science and Technology,
Department of Electronic Engineering,
Douliou, 64002 Yunlin, Taiwan
{huangchs,M10213327}@yuntech.edu.tw

Abstract. In this study, the effects of four indoor illuminations on e-book readers are investigated by HRV analysis. Two main types of commercial light bulbs are adopted, the saving energy and LED ones. They are all categorized as high color temperature and low color temperature. The results of HRV analysis indicate that subjects feel more awake under high color temperature illuminations, no matter the type of bulbs. Moreover, subjects perform better in Continuous Performance Test when high color temperature energy-saving bulb is used as illumination source.

Keywords: indoor illumination, e-book, HRV.

1 Introduction

Color is the manifestation of light by means of the reflection over the object. Human's feeling are affected not only by luminance, but also by colors[1]. The indoor illumination provides the basic light environment while we are in the house. The color temperature is the index of commercial light bulb which tells us the wavelength spectrum of the output light. For high color temperature, called cool light, the blue color is dominant, and for low color temperature, called warm light, the red one is dominant. As green technology emerging, the energy saving is one of the issue, and traditional incandescent light bulbs have been replaced by energy saving bulbs or LED bulbs.

As people nowadays are pursuing healthy life and working efficiency, the mobile devices penetrate our daily life. Many people use the mobile device to read messages, news, and even books. We may call these behaviors e-book reading. The reading experience is different from the traditional paper reading. Since most of the mobile devices adopt active lighting, the indoor illumination becomes more and more important to protect human's eyes to prevent glaring. Therefore, while choosing the bulb, we should not only consider the energy conversion efficiency and color rendering, but also the human factors to establish a comfortable environment in any case[2].

* Corresponding author.

H. Sun et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 329, DOI: 10.1007/978-3-319-12286-1_22

In this work, four types of bulbs as experimental illumination: energy-saving bulbs with high color temperature (6500 K) and low color temperature (2700 K), and LED bulbs with high color temperature (6500 K) and low color temperature (2700 K). For convenience, we shorten them as WS, YS, WL, YL, respectively.

2 Experimental Design

20 healthy male were participated in this experiment, each subject accepted four light illumination groups: cool white energy-saving bulb (WS), yellow white energy-saving bulb (YS), cool white LED bulb (WL) and yellow white LED bulb (YL). A questionnaire was then conducted in which each participant was requested to list their subjective feeling after the experiment.

As the experiment proceeds, subjects were first rest for 7 minutes in slightly natural light, than reading e-book under selected illumination for 14 minutes. And next Continuous Performance Test (CPT) would be conducted on display screen. The ECG signals had been recorded all time.

The whole experimental timescale are shown in figure 1.

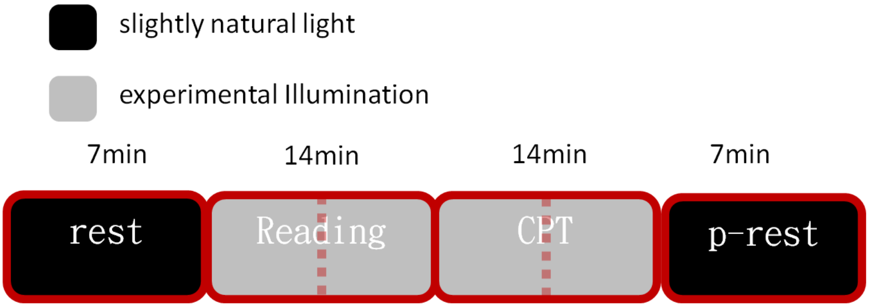


Fig. 1. Experimental Timescale

3 Data Analysis

Figure 2~4 showed the HRV analysis of normalized low frequency power (nLFP), normalized high frequency power (nHFP), low frequency power to high frequency power ratio (L/H ratio). All the 14 minutes reading is divided into 2 periods, the first 7 minutes reading and the last 7 minutes reading. As in Figure 3, we see there are statistic significant differences for the first 7 minutes reading test in four groups, but only the two cool white illumination groups (WS, WL) also show statistic significant differences in the last 7 minutes reading. Figure 4 is HRV L/H ratio, and only the WS group shows statistic significant difference.

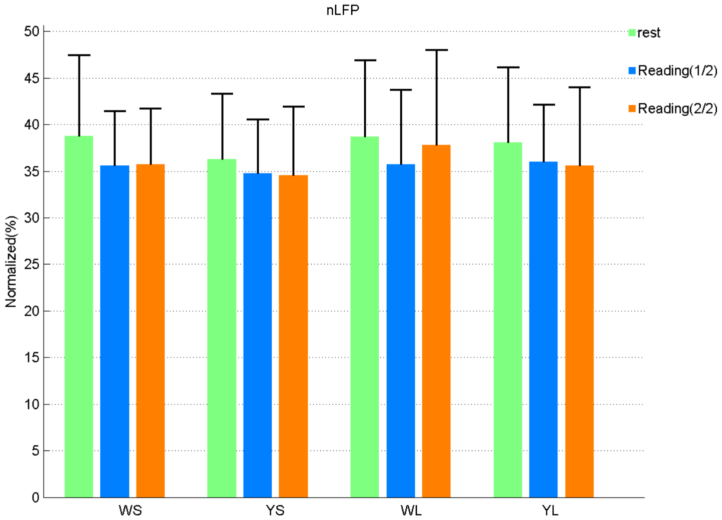


Fig. 2. nLFP Analysis

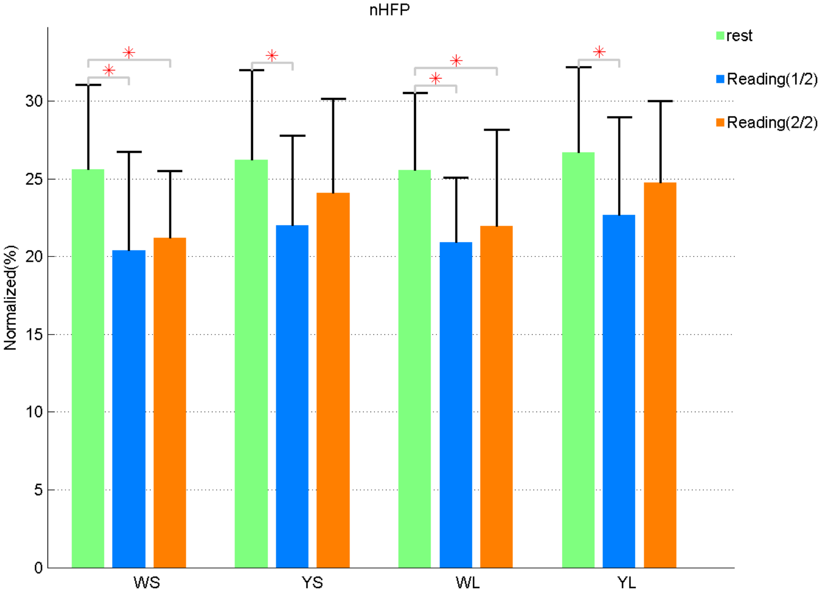


Fig. 3. nHFP Analysis

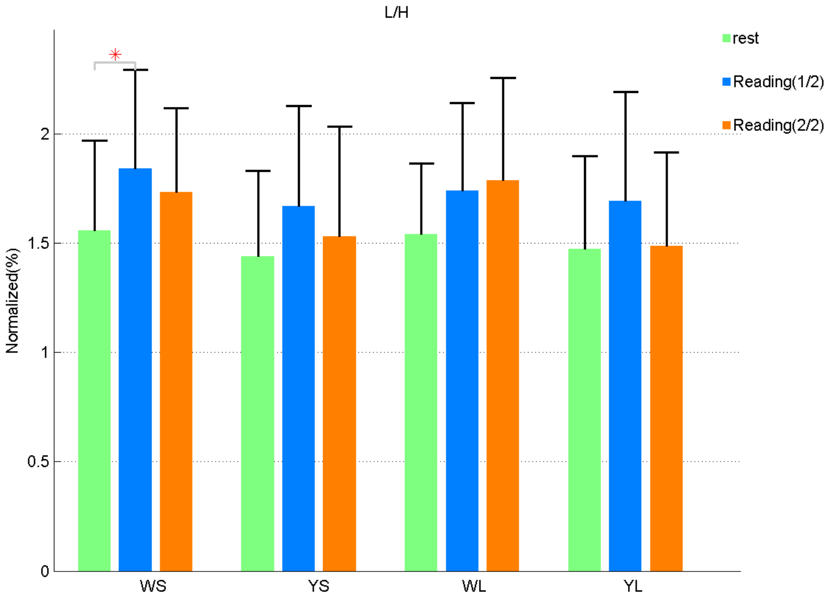


Fig. 4. L/H Analysis

Figures 5 and 6 show the omission error rate and commission error rate in CPT test, respectively. We can see that the WS group have lowest error rate in both, and the YL groups have the highest error rate in both combination.

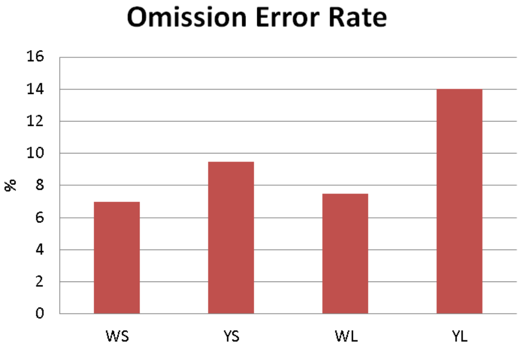


Fig. 5. Omission Error Rate in CPT

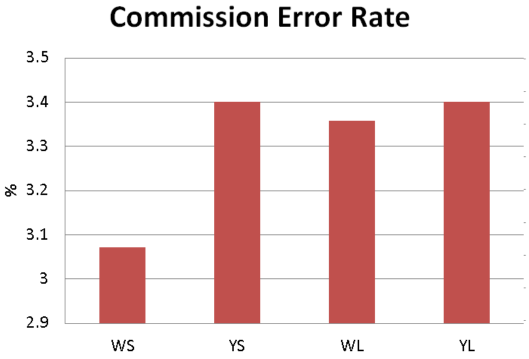


Fig. 6. Commission Error Rate in CPT

4 Discussion

nHFP is corresponding to sympathetic nervous system. It can also be the index of relaxation level[3][4]. In figure 3, it is obvious that nHFP of the two yellow white light groups increases in the last 7 minutes reading. We assume yellow white light created a more relaxing environment that occur drowsiness to the subject, so that the nHFP increased.

In CPT analysis, WS group performed best in both omission rate (7.0%) and commission error rate (3.1%), indicating their highest concentration level[5]. And YL groups have highest error rate in both combination (14% and 3.4%). These seemed to be correlated to L/H ration.

5 Conclusion

In this study, tests of e-book reading under four light bulbs as indoor illumination were conducted. As the result shown, the cool white bulbs, no matter energy saving or LED types, may keep subjects awake while they are reading for the whole 14 minutes. And the CPT test reveals some further issues for study.

References

1. Yoto, A., Katsuura, T., Iwanaga, K., Shimomura, Y.: Effects of Object Color Stimuli on Human Brain Activities in Perception and Attention Referred to EEG alpha Band Response. J. Physiol. Anthropol. 26, 373–379 (2007)
2. Aquirre, R.C., Colombo, E.M., Barraza, J.F.: Effect of glare on simple reaction time. Optical Society of America 25(7), 1790–1798 (2008); Foster, I., Kesselman, C.: The Grid: Blueprint for a New Computing Infrastructure. Morgan Kaufmann, San Francisco (1999)

3. Matsumura, K., Miura, K., Takata, Y., Kurokawa, H., Kajiyama, M., Abe, I., Fujishima, M.: Changes in blood pressure and heart rate variability during dental surgery. American Journal of Hypertension 11, 1376–1380 (1998); Foster, I., Kesselman, C., Nick, J., Tuecke, S.: The Physiology of the Grid: an Open Grid Services Architecture for Distributed Systems Integration. Technical report, Global Grid Forum (2002)
4. Minami, J., Yoshii, M., Todoroki, M., Nishikimi, T., Ishimitsu, T., Fukunaga, T., Matsuoka, H.: Effects of alcohol restriction on ambulatory blood pressure, heart rate, and heart rate variability in Japanese men. American Journal of Hypertension 15, 125–129 (2002)
5. Riccio, C.A., Reynolds, C.R., Lowe, P., Moore, J.J.: The continuous performance test: a window on the neural substrates for attention? Arch. Clin. Neuropsychol. 17, 235–272 (2002)

Part IV
Image Processing and Intelligent Applications

Modified Choice Function Heuristic Selection for the Multidimensional Knapsack Problem

John H. Drake[1], Ender Özcan[1], and Edmund K. Burke[2]

[1] ASAP Research Group,
School of Computer Science, University of Nottingham,
Jubilee Campus, Wollaton Road, Nottingham, NG8 1BB, UK
{psxjd2,ender.ozcan}@nottingham.ac.uk
[2] Computing Science and Mathematics, School of Natural Sciences
University of Stirling, Stirling, FK9 4LA, Scotland
e.k.burke@stir.ac.uk

Abstract. Hyper-heuristics are a class of high-level search methods used to solve computationally difficult problems, which operate on a search space of low-level heuristics rather than solutions directly. Previous work has shown that selection hyper-heuristics are able to solve many combinatorial optimisation problems, including the multidimensional 0-1 knapsack problem (MKP). The traditional framework for iterative selection hyper-heuristics relies on two key components, a heuristic selection method and a move acceptance criterion. Existing work has shown that a hyper-heuristic using *Modified Choice Function* heuristic selection can be effective at solving problems in multiple problem domains. *Late Acceptance Strategy* is a hill climbing metaheuristic strategy often used as a move acceptance criteria in selection hyper-heuristics. This work compares a *Modified Choice Function - Late Acceptance Strategy* hyper-heuristic to an existing selection hyper-heuristic method from the literature which has previously performed well on standard MKP benchmarks.

Keywords: Hyper-heuristics, Choice Function, Heuristic Selection, Multidimensional Knapsack Problem, Combinatorial Optimization

1 Introduction

Hyper-heuristics are high-level search methodologies which operate on a search space of heuristics. A hyper-heuristic is defined by Burke et al. [1,2] as: '...a search method or learning mechanism for selecting or generating heuristics to solve computational search problems'. This definition covers the two main classes of hyper-heuristics, those concerned with heuristic selection and those with heuristic generation. Although often considered as an alternative to metaheuristics, the recent definition of a metaheuristic offered by Sörensen and Glover [3] somewhat subsumes hyper-heuristics. According to their definition, a selection hyper-heuristic is a metaheuristic which provides a framework within which to mix

H. Sun et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 329, DOI: 10.1007/978-3-319-12286-1_23

and control low-level heuristics, whereas a generation hyper-heuristic is a metaheuristic to generate heuristics. It follows that if a metaheuristic is a heuristic, a hyper-heuristic can either be a metaheuristic itself (e.g. a Grammatical Evolution system to generate heuristics [4]) or contain metaheuristic components (e.g. a selection hyper-heuristic using *Late Acceptance Strategy* move acceptance criterion [5]). Hyper-heuristics have been applied successfully to a wide range of problems including: production scheduling [6], nurse rostering [7], examination timetabling [7], sports scheduling [8], bin packing [9], dynamic environments [10], vehicle routing [4] and the multidimensional 0-1 knapsack problem [11,12].

A traditional selection hyper-heuristic iteratively selects and applies low-level heuristics to a single solution, with a decision made at each step whether to accept the new solution. In this paper, such hyper-heuristics are labelled *heuristic selection method - move acceptance criterion* hereafter. Özcan et al. [13] described four different hyper-heuristic frameworks. One of these frameworks F_C, selects and applies a mutational heuristic from a set of low-level heuristics, followed by a pre-defined hill climber before deciding whether to accept the new solution. This is the framework used in this paper, illustrated in Figure 1.

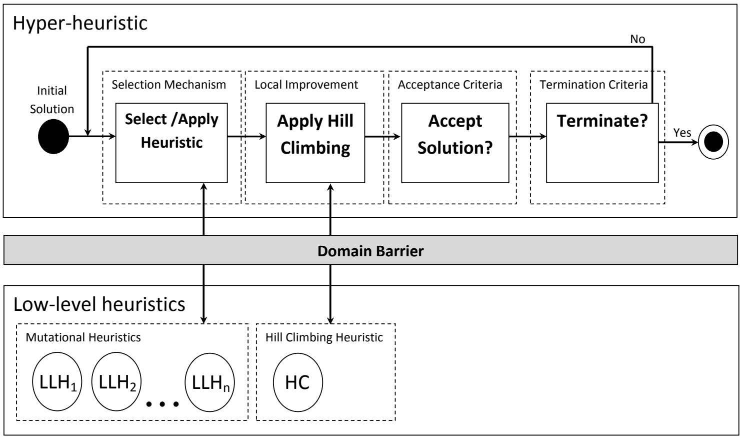


Fig. 1. F_C single-point search hyper-heuristic framework

The *Modified Choice Function* is an elegant heuristic selection method inspired by Reinforcement Learning, which scores heuristics based on a combination of three different measures. At each given step of a search, the heuristic selected is based on a weighted combination of these scores. Existing work has shown that a hyper-heuristic using *Modified Choice Function* heuristic selection can be effective at solving problems in multiple problem domains. In this paper we investigate the suitability of using *Modified Choice Function* hyper-heuristics to solve the MKP, a domain in which hyper-heuristics have previously shown to be successful.

2 Selection Hyper-Heuristics and *Choice Function* Heuristic Selection

The first use of the term hyper-heuristic was by Cowling et al. [14], who defined hyper-heuristics as '*heuristics to choose heuristics*'. This paper investigated the application of a number of *Simple Random*, *Greedy* and *Choice Function*-based hyper-heuristic approaches to a real-world sales summit scheduling problem using two deterministic move acceptance criteria: *All Moves* and *Only Improving. Choice Function* heuristic selection has also been used by Bilgin et al. [15] for benchmark function optimisation, Özcan et al. [16] and Burke et al. [17] for examination timetabling and Drake et al. [12] for the MKP.

The *Choice Function* is a heuristic selection method which scores heuristics based on a combination of three different measures. The first measure (f_1) reflects the previous performance of a given low-level heuristic, weighted towards the most recent application. The value of f_1 for a low-level heuristic h_j is defined as:

$$f_1(h_j) = \sum_n \alpha^{n-1} \frac{I_n(h_j)}{T_n(h_j)} \tag{1}$$

where $I_n(h_j)$ is the change in solution quality, $T_n(h_j)$ is the time taken to execute the heuristic for each previous invocation n of heuristic h_j and α is a value between 0 and 1 giving greater importance to recent performance.

The second measure (f_2) attempts to capture any pair-wise dependencies between heuristics. Values of f_2 are calculated for each heuristic h_j when invoked immediately following h_k using the formula in Equation 6:

$$f_2(h_k, h_j) = \sum_n \beta^{n-1} \frac{I_n(h_k, h_j)}{T_n(h_k, h_j)} \tag{2}$$

where $I_n(h_k, h_j)$ is the change in evaluation function, $T_n(h_k, h_j)$ is the time taken to call the heuristic for each previous invocation n of heuristic h_j following h_k and β is a value between 0 and 1 which also gives greater importance to recent performance.

The third measure (f_3) is the time elapsed ($\tau(h_j)$) since the heuristic was last selected by the *Choice Function.* This allows all heuristics at least a small chance of selection.

$$f_3(h_j) = \tau(h_j) \tag{3}$$

In order to rank heuristics, a score is given to each heuristic with *Choice Function F* calculated as:

$$F(h_j) = \alpha f_1(h_j) + \beta f_2(h_k, h_j) + \delta f_3(h_j) \tag{4}$$

where α and β as defined previously weight f_1 and f_2 respectively to provide intensification of the heuristic search process whilst δ weights f_3 to provide sufficient diversification.

2.1 *Modified Choice Function* Heuristic Selection

The *Modified Choice Function* [18] was introduced to overcome some of the limitations of the classic *Choice Function*, when applied to cross-domain optimisation using the CHeSC2011 [19] benchmarks. *Modified Choice Function* heuristic selection automatically controls the intensification and diversification parameters of the *Choice Function*, weighting each of f_1, f_2 and f_3 using a method inspired by Reinforcement Learning, giving far greater emphasis to intensification. The *Modified Choice Function* does not make a distinction between the values of α or β which weight f_1 and f_2, so these are reduced to a single intensification parameter which is referred to as ϕ. Using the *Modified Choice Function*, the score F_t for each heuristic h_j is defined as:

$$F_t(h_j) = \phi_t f_1(h_j) + \phi_t f_2(h_k, h_j) + \delta_t f_3(h_j) \tag{5}$$

where t is the current iteration. At each stage if an improvement in solution quality is observed ϕ is rewarded heavily whilst δ is harshly punished. If the quality of the solution does not improve following the application of a low-level heuristic, the level of intensification is decreased linearly, reducing ϕ and increasing δ to diversify the heuristic search process. The intention is to give the intensification component of the *Choice Function* more time as the dominating factor in the calculation of F. In this paper the parameters ϕ_t and δ_t are defined as:

$$\phi_t = \begin{cases} 0.99, & \text{if an improving move is made} \\ \max\{\phi_{t-1} - 0.01, 0.01\}, & \text{if a non-improving move is made} \end{cases} \tag{6}$$

$$\delta_t = 1 - \phi_t \tag{7}$$

The use of 0.01 as the minimum value ensures that the diversification component of the *Modified Choice Function* always has some non-negative influence on the F value for each heuristic. Here we would like to clarify that whilst each individual heuristic has its own F value, all low low level heuristics use the same ϕ and δ values. The *Modified Choice Function* was shown to outperform the original *Choice Function* on average, over six different problem domains by Drake et al. [18]. Additionally the best results in the literature were achieved for the MAX-SAT problem domain, within the HyFlex framework.

3 The Multidimensional Knapsack Problem

The NP-hard [20] multidimensional 0-1 knapsack problem (MKP) [21] is a generalised case of the standard 0-1 knapsack problem, with roots in applications such as capital budgeting and project selection. The MKP is a resource allocation model, where the objective is to select a subset of objects which yield the greatest profit, whilst observing the constraints on knapsack capacities. Unlike the standard 0-1 knapsack problem, each object j consumes a different amount of resources in each dimension i when selected. Formally the MKP is stated as:

$$\text{maximise} \qquad \sum_{j=1}^{n} p_j x_j \tag{8}$$

$$\text{subject to} \qquad \sum_{j=1}^{n} a_{ij} x_j \leq b_i, \qquad i = 1, ..., m \tag{9}$$

$$\text{with} \qquad x_j \in \{0, 1\}, \qquad j = 1, ..., n \tag{10}$$

where p_j is the profit for selecting item j, a_{ij} is the resource consumption of item j in dimension i, b_i is the capacity constraint of each dimension i. Using direct binary encoding, x_1,...,x_n is a set of decision variables indicating whether or not object j is included in the knapsack. The size of a problem is defined by the total number of variables n and the number of dimensions m.

A number of benchmarks sets exist for the MKP, each with different properties[1]. SAC-94 is a benchmark library of MKP instances taken from a number of papers in the literature, often representing real-world examples. These instances are generally small with m ranging from 2 to 30 and n ranging from 10 to 105 with optimal solutions known for all. ORLib is a widely used benchmark library containing 270 instances, split into 27 sets of 10 instances, containing $n \in \{100, 250, 500\}$ variables, $m \in \{5, 10, 30\}$ dimensions and *tightness ratio* $\in \{0.25, 0.50, 0.75\}$. A third benchmark library was introduced by Glover and Kochenbeger [22], referred to here as GK, containing much larger instances of the MKP with n up to 2500 and m up to 100. As optimal solutions are not known for all of the instances in ORLib and GK, performance is often measured using the %-gap distance to the solution to the LP-relaxed problem calculated as:

$$100 * \frac{LPopt - SolutionFound}{LPopt} \tag{11}$$

where *LPopt* is the fitness value of the LP-relaxed solution to a given problem and *SolutionFound* is the fitness value of the solution found. In the case of the SAC-94 instances, as optimal solutions are known, performance is measured by the proportion of instances for which the optimal solution is found.

4 Experimental Framework and Parameters

The *Modified Choice Function* described in the previous section is paired with *Late Acceptance Strategy* acceptance criterion and applied to the MKP benchmarks. *Late Acceptance Strategy* [23] is a general purpose optimisation technique which has previously been used as an acceptance criterion in a number of selection hyper-heuristics [16,11,5]. A *Choice Function* - *Late Acceptance Strategy*

[1] All three benchmark instance sets have been standardised and are available in a unified format at: `http://www.cs.nott.ac.uk/~jqd/mkp/index.html`

hyper-heuristic was shown to be the best of nine selection hyper-heuristics for the MKP by Drake et al. [11]. The same framework is used here, containing seven low-level heuristics to select from. A single hill climbing heuristic to be applied after each low-level, as required by an F_C hyper-heuristic framework, is also included. In each case a run for a single instance terminates once 10^6 fitness evaluations have been performed. This allows for direct comparison to the results of the original *Choice Function - Late Acceptance Strategy* hyper-heuristic and a wide range of existing methods in the literature. A single run of each hyper-heuristic is performed on each of the instances in the ORLib and SAC-94 benchmark sets. In the case of the larger GK instances, results are given as the average of 10 runs for each instance.

5 Computational Results

Table 1 shows the results of the *Modified Choice Function - Late Acceptance Strategy* hyper-heuristic over the 270 ORLib instances in terms of average %-gap, along with the results for *Choice Function - Late Acceptance Strategy.* Standard deviations are given as subscript. The results for both hyper-heuristics are very similar over the ORLib instances, with *Choice Function - Late Acceptance Strategy* obtaining an average %-gap of 0.70 and *Modified Choice Function - Late Acceptance Strategy* a %-gap of 0.73, with little difference in standard deviation. An independent Student's t-test within a 95% confidence interval shows no statistically significant difference between the two hyper-heuristics on this benchmark set.

Table 2(a) and Table 2(b) compare the performance of *Choice Function - Late Acceptance Strategy* and *Modified Choice Function - Late Acceptance Strategy* over the SAC-94 and GK problem instances. In the case of the SAC-94 instances, which contains six subsets of problems, the success rate is defined as the proportion on instances within each subset that the optimal solution is found. Again the two methods are showing very similar performance on both of these benchmark sets. *Modified Choice Function - Late Acceptance Strategy* is slightly outperformed in terms of success rate in the pb and weish instances, finding the optimal solution in one and three less instances respectively compared to *Choice Function - Late Acceptance Strategy.* On the GK benchmark set, both methods obtain the same average %-gap over 10 runs of each of the 11 instances.

Table 3 gives the average %-gap for a number of techniques from the literature over the ORLib benchmark set. Our approach is able to outperform many of the existing methods, achieving an average %-gap of 0.73. This is considerably lower than many of the heuristic methods an some of the metaheuristic techniques proposed previously.

Table 1. Comparison between *Choice Function - Late Acceptance Strategy* (CF-LAS) and *Modified Choice Function - Late Acceptance Strategy* (MCF-LAS) on all 270 instances of ORLib in terms of %-gap

Instance Set	CF-LAS	MCF-LAS
OR5x100-0.25	$1.16_{0.20}$	$1.09_{0.21}$
OR5x100-0.50	$0.53_{0.08}$	$0.57_{0.08}$
OR5x100-0.75	$0.40_{0.07}$	$0.38_{0.05}$
OR5x250-0.25	$0.42_{0.04}$	$0.41_{0.10}$
OR5x250-0.50	$0.20_{0.03}$	$0.22_{0.04}$
OR5x250-0.75	$0.13_{0.01}$	$0.14_{0.02}$
OR5x500-0.25	$0.19_{0.03}$	$0.21_{0.04}$
OR5x500-0.50	$0.10_{0.03}$	$0.10_{0.03}$
OR5x500-0.75	$0.06_{0.01}$	$0.06_{0.01}$
OR10x100-0.25	$2.00_{0.22}$	$1.87_{0.16}$
OR10x100-0.50	$1.02_{0.19}$	$0.95_{0.16}$
OR10x100-0.75	$0.58_{0.08}$	$0.53_{0.09}$
OR10x250-0.25	$0.83_{0.09}$	$0.79_{0.11}$
OR10x250-0.50	$0.39_{0.06}$	$0.41_{0.05}$
OR10x250-0.75	$0.23_{0.03}$	$0.24_{0.03}$
OR10x500-0.25	$0.40_{0.06}$	$0.44_{0.07}$
OR10x500-0.50	$0.18_{0.02}$	$0.20_{0.03}$
OR10x500-0.75	$0.12_{0.01}$	$0.13_{0.01}$
OR30x100-0.25	$3.45_{0.46}$	$3.61_{0.53}$
OR30x100-0.50	$1.56_{0.26}$	$1.60_{0.29}$
OR30x100-0.75	$0.92_{0.08}$	$0.97_{0.15}$
OR30x250-0.25	$1.55_{0.17}$	$1.75_{0.22}$
OR30x250-0.50	$0.71_{0.08}$	$0.79_{0.10}$
OR30x250-0.75	$0.39_{0.04}$	$0.43_{0.07}$
OR30x500-0.25	$0.92_{0.10}$	$1.05_{0.10}$
OR30x500-0.50	$0.39_{0.05}$	$0.44_{0.06}$
OR30x500-0.75	$0.23_{0.02}$	$0.27_{0.02}$
Average$_{StdDev}$	$0.70_{0.09}$	$0.73_{0.11}$

Table 2. Performance of *Choice Function - Late Acceptance Strategy* and *Modified Choice Function - Late Acceptance Strategy* in terms of (a) success rate of over SAC-94 instances and (b) %-gap obtained in GK instances

(a)

Dataset	CF-LAS	MCF-LAS
hp	0.00	0.00
pb	0.67	0.50
pet	0.50	0.50
sento	1.00	1.00
weing	0.63	0.63
weish	1.00	0.90

(b)

Instance	CF-LAS	MCF-LAS
GK01	0.57 $_{1.49}$	0.58 $_{1.66}$
GK02	0.81 $_{3.86}$	0.78 $_{3.75}$
GK03	0.63 $_{3.10}$	0.63 $_{4.30}$
GK04	0.91 $_{3.77}$	0.86 $_{5.81}$
GK05	0.45 $_{3.00}$	0.44 $_{5.83}$
GK06	0.76 $_{5.02}$	0.78 $_{4.04}$
GK07	0.19 $_{6.48}$	0.22 $_{2.88}$
GK08	0.33 $_{5.68}$	0.31 $_{7.60}$
GK09	0.07 $_{7.47}$	0.07 $_{6.85}$
GK10	0.14 $_{8.68}$	0.14 $_{11.71}$
GK11	0.13 $_{12.34}$	0.14 $_{11.10}$
Average	0.45$_{0.30}$	0.45$_{0.29}$

Table 3. Comparison of genetic programming hyper-heuristic to previous approaches over all instances in ORLib in terms of %-gap

Type	Reference	%-gap
MIP	Drake et al. [11] (CPLEX 12.2)	0.52
MA	Chu and Beasley [24]	0.54
Selection HH	Drake et al. [11]	0.70
Selection HH	***Modified Choice Function - Late Acceptance Strategy***	**0.73**
MA	Özcan and Basaran [25]	0.92
Heuristic	Pirkul [26]	1.37
Heuristic	Fréville and Plateau [27]	1.91
Generation HH	Drake et al. [12]	3.04
MIP	Chu and Beasley [24] (CPLEX 4.0)	3.14
Heuristic	Akçay et al. [28]	3.46
Heuristic	Volgenant and Zoon [29]	6.98
Heuristic	Magazine and Oguz [30]	7.69

6 Conclusions

In this work we have applied a *Modified Choice Function - Late Acceptance Strategy* selection hyper-heuristic to a well known optimisation problem, the MKP. Previously, using *Modified Choice Function* heuristic selection was shown to outperform the classic *Choice Function.* Additionally the *Choice Function* has previously worked well with *Late Acceptance Strategy* move acceptance when solving the MKP. Although the *Modified Choice Function* has outperformed the *Choice Function* over multiple domains in the past, this has not been the case when applied to the MKP, with the *Modified Choice Function* offering slightly poorer performance in two of the three benchmark sets tested. Future work will combine the *Modified Choice Function* with other move acceptance criteria, to assess whether there is any variation in performance over the MKP benchmarks. We will also test the *Modified Choice Function* on further benchmark problems in order to better understand the type of problem in which this heuristic selection method can perform well.

References

1. Burke, E.K., Hyde, M., Kendall, G., Ochoa, G., Özcan, E., Woodward, J.: A Classification of Hyper-heuristic Approaches. In: Handbook of Metaheuristics, 2nd edn., pp. 449–468. Springer (2010)
2. Burke, E.K., Hyde, M., Kendall, G., Ochoa, G., Özcan, E., Qu, R.: Hyper-heuristics: A survey of the state of the art. Journal of the Operational Research Society 64(12), 1695–1724 (2013)
3. Sörensen, K., Glover, F.: Metaheuristics. In: Encyclopedia of Operations Research and Management Science, pp. 960–970. Springer (2013)
4. Drake, J.H., Kililis, N., Özcan, E.: Generation of vns components with grammatical evolution for vehicle routing. In: Krawiec, K., Moraglio, A., Hu, T., Etaner-Uyar, A.Ş., Hu, B. (eds.) EuroGP 2013. LNCS, vol. 7831, pp. 25–36. Springer, Heidelberg (2013)
5. Jackson, W.G., Özcan, E., Drake, J.H.: Late acceptance-based selection hyper-heuristics for cross-domain heuristic search. In: Proceedings of the 13th Annual Workshop on Computational Intelligence (UKCI 2013), pp. 228–235. IEEE Press, Surrey (2013)
6. Fisher, H., Thompson, G.: Probabilistic learning combinations of local job-shop scheduling rules. In: Factory Scheduling Conference, Carnegie Institute of Technology (1961)
7. Burke, E.K., Kendall, G., Soubeiga, E.: A tabu-search hyperheuristic for timetabling and rostering. Journal of Heuristics 9(6), 451–470 (2003)
8. Gibbs, J., Kendall, G., Özcan, E.: Scheduling english football fixtures over the holiday period using hyper-heuristics. In: Schaefer, R., Cotta, C., Kołodziej, J., Rudolph, G. (eds.) PPSN XI, Part I. LNCS, vol. 6238, pp. 496–505. Springer, Heidelberg (2010)
9. López-Camacho, E., Terashima-Marín, H., Ross, P.: A hyper-heuristic for solving one and two-dimensional bin packing problems. In: Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2011), pp. 257–258. ACM, Dublin (2011)
10. Kiraz, B., Uyar, A.S., Özcan, E.: Selection hyper-heuristics in dynamic environments. Journal of the Operational Research Society 64(12), 1753–1769 (2013)
11. Drake, J.H., Özcan, E., Burke, E.K.: Controlling crossover in a selection hyper-heuristic framework. Technical Report No. NOTTCS-TR-SUB-1104181638-4244, School of Computer Science, University of Nottingham (2011)
12. Drake, J.H., Hyde, M., Ibrahim, K., Özcan, E.: A genetic programming hyper-heuristic for the multidimensional knapsack problem. In: Proceedings of the 11th IEEE International Conference on Cybernetic Intelligent Systems (CIS 2012), pp. 76–80. IEEE Press, Limerick (2012)

13. Özcan, E., Bilgin, B., Korkmaz, E.E.: A comprehensive analysis of hyper-heuristics. Intelligent Data Analysis 12(1), 3–23 (2008)
14. Cowling, P., Kendall, G., Soubeiga, E.: A hyperheuristic approach to scheduling a sales summit. In: Burke, E., Erben, W. (eds.) PATAT 2000. LNCS, vol. 2079, pp. 176–190. Springer, Heidelberg (2001)
15. Bilgin, B., Özcan, E., Korkmaz, E.E.: An experimental study on hyper-heuristics and exam timetabling. In: Burke, E.K., Rudová, H. (eds.) PATAT 2007. LNCS, vol. 3867, pp. 394–412. Springer, Heidelberg (2007)
16. Özcan, E., Bykov, Y., Birben, M., Burke, E.K.: Examination timetabling using late acceptance hyper-heuristics. In: Proceedings of the IEEE Congress on Evolutionary Computation (CEC 2009), pp. 997–1004. IEEE Press, Trondheim (2009)
17. Burke, E.K., Kendall, G., Misir, M., Özcan, E.: Monte carlo hyper-heuristics for examination timetabling. Annals of Operations Research 196(1), 73–90 (2012)
18. Drake, J.H., Özcan, E., Burke, E.K.: An improved choice function heuristic selection for cross domain heuristic search. In: Coello, C.A.C., Cutello, V., Deb, K., Forrest, S., Nicosia, G., Pavone, M. (eds.) PPSN 2012, Part II. LNCS, vol. 7492, pp. 307–316. Springer, Heidelberg (2012)
19. Ochoa, G., Hyde, M.: The cross-domain heuristic search challenge (CHeSC 2011) (2011), http://www.asap.cs.nott.ac.uk/chesc2011/
20. Garey, M.R., Johnson, D.S.: Computers and Intractability: A Guide to the Theory of NP-Completeness. Freeman & Co., New York (1979)
21. Weingartner, H.M., Ness, D.N.: Methods for the solution of the multidimensional 0/1 knapsack problem. Operations Research 15(1), 83–103 (1967)
22. Glover, F., Kochenberger, G.: Benchmarks for "the multiple knapsack problem" (n.d.), http://hces.bus.olemiss.edu/tools.html
23. Burke, E.K., Bykov, Y.: A late acceptance strategy in hill-climbing for exam timetabling problems. In: Proceedings of the International Conference on the Practice and Theory of Automated Timetabling (PATAT 2008), Montreal, Canada (2008) Extended Abstract
24. Chu, P.C., Beasley, J.E.: A genetic algorithm for the multidimensional knapsack problem. Journal of Heuristics 4(1), 63–86 (1998)
25. Özcan, E., Basaran, C.: A case study of memetic algorithms for constraint optimization. Soft Computing 13(8-9), 871–882 (2009)
26. Pirkul, H.: A heuristic solution procedure for the multiconstraint zero-one knapsack problem. Naval Research Logistics 34(2), 161–172 (1987)
27. Freville, A., Plateau, G.: An efficient preprocessing procedure for the multidimensional 0-1 knapsack problem. Discrete Applied Mathematics 49(1-3), 189–212 (1994)
28. Akçay, Y., Li, H., Xu, S.H.: Greedy algorithm for the general multidimensional knapsack problem. Annals of Operations Research 150(1), 17–29 (2007)
29. Volgenant, A., Zoon, J.A.: An improved heuristic for multidimensional 0-1 knapsack problems. Journal of the Operational Research Society 41(1), 963–970 (1990)
30. Magazine, M.J., Oguz, O.: A heuristic algorithm for the multidimensional zero-one knapsack problem. European Journal of Operational Research 16(3), 319–326 (1984)

A Distributed Computational Model of State of Consciousness

Susmit Bagchi

Department of Informatics, Gyeongsang National University,
Jinju, South Korea 660 701
susmitbagchi@yahoo.co.uk

Abstract. The computational modeling of neurobiological phenomena is an important area in bio-inspired computing. The understanding of state of consciousness as cognitive function is central to it. The functioning of neurological structures is inherently distributed in nature having a closer match to distributed computing. However, the role of functional neurophysiology is critical in cognition modeling. This paper proposes a mathematical model of state of consciousness by mapping the functional neurophysiology and by inducing distributed computing structures in coherence. The scopes of evolution of consciousness and memory are incorporated into the model. The numerical simulation of the distributed computational model is conducted by considering different choice functions. The results illustrate that, gradual evolution of positive consciousness is deterministic under fair excitation from environment.

Keywords: Cognition, consciousness, bio-inspired computing, distributed computing.

1 Introduction

The consciousness is a neurobiological phenomenon of a living being. At the physical layer of consciousness, a biological system interacts with the environment through the sensors and processes the input signals in brain to generate output. However, the modular cerebral network of brain often processes the inputs in the unconscious state [7]. On the other hand, the conscious and abstract thinking happens at the global workspace (GW) level [2, 7]. Several computational models are proposed to understand consciousness following artificial neural network and probabilistic reasoning. In general, the attributes of defining consciousness are vague and incomplete, which have resulted in incompleteness of computational models of consciousness [8]. The main reason is that, the models aim at human-centric consciousness definitions rather than trying to create a functional model of consciousness following neurobiological phenomena of living species. The understanding and modeling of cognitive actions and consciousness require the quantitative and theoretical frameworks bridging the neurobiological functions and the computational as well as algorithmic functions [1]. The cognitive functions of brain generating consciousness are inherently a distributed computing mechanism,

H. Sun et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 329, DOI: 10.1007/978-3-319-12286-1_24

because the information processing happens at different locations in brain due to an input [7]. This paper proposes a novel computational model of deterministic consciousness following the functional attributes of neurobiological phenomena. The proposed model combines the distributed computing concepts to the functional attributes of neurophysiology to understand the state of deterministic consciousness. The concepts of evolution and memory are incorporated into the model to understand the dynamics of state of deterministic consciousness. The experimental evaluations are conducted by numerical computation aiming to quantify deterministic consciousness and to follow its dynamics. The proposed model and the results illustrate that, consciousness can be possibly quantified and the dynamical states can be traced if appropriate choice functions are selected matching neurobiological phenomena. The rest of the paper is organized as follows. Section 2 outlines related work. Section 3 and 4 describe the mathematical model of state of consciousness and experimental evaluation, respectively. Section 5 concludes the paper.

2 Related Work

The physiological level of understanding of cognitive capabilities involves neuronal structures and signal transductions. The neural correlation models are proposed by researchers to explain the development of consciousness in brain [2, 11, 13]. The mind-brain relationship and structural analysis are derived through experimentation involving EEG, fMRI and, PET [1, 2]. However, the physiological understanding of brain structures cannot explain the cognitive capabilities and consciousness in complete form. It has been proposed that neurological cognitive actions can be explained by following computational models such as, finite state automation as well as push-down stack [1, 9, 10]. Following this approach, a single neuron is modeled as tree-shaped computing structure [1]. The tree-model tries to map the physiological structure and functions of neurons into the computational structure. At the functional level, the predictive coding model of neocortex is proposed [12, 14]. The predictive coding model is a generalized model and incorporates elements of distributed computation. It is observed that, at the algorithmic levels, the cognitive functions become computationally intractable although living species perform such functions without delay [15]. The expression of intension is a cognitive function too having neurological mechanisms. Researchers have proposed models to predict intention based on pattern matching and sequence analysis [5]. The simulation theory is proposed by researchers to explain the intention prediction [18]. According to simulation theory, cross-simulation and cross-observation between two subjects can be performed making one subject aware of intension of another by using own state of cognition and consciousness. However, this model is based on empiricism.

In general, the existing models of cognitive functions and consciousness can be broadly classified into six categories [2, 4, 6], namely: (1) global-workspace model, (2) information integration model, (3) internal self-model, (4) higher-level representation theory, (5) theory of attention mechanisms and, (6) virtual machine formalism. The global-workspace model considers interconnection of distributed cerebral networks and its activation for certain duration of time. The consciousness model based on artificial neural network (ANN) is proposed following global-workspace concept representing abstract thinking [7]. On the other hand, the cognition capability of abstract thinking is modeled by Central-Representation architecture along with monitor system [16]. However, these models are complex, computationally expensive and do not consider evolutionary consciousness. It is observed that, neurological phenomena of consciousness and experiences are not instantaneous neuronal activities and, the retention (memory) has a role into it [3]. Furthermore, the expression of experiences has a computational basis, which can be modeled by mathematical frameworks [17, 19, 20].

3 Model of Consciousness and State Dynamics

From the physiological as well as anatomical point of view, different sections of a brain are responsible to process different excitations from environment. Experimentations have revealed that, the information processing in brain is inherently distributed in nature. However, there is a mechanism of coherency and coordination among the neuronal sections to produce a conscious output. In view of computation, a brain can be modeled as a tightly-coupled distributed computing system, where specialized neuronal sections are represented as the nodes connected to each other through neuro-network. In this paper, a map the brain-model in view of distributed computing structures is formulated and the computational model of state of consciousness is constructed following functional neurophysiology. The computational view of brain is illustrated in Fig. 1.

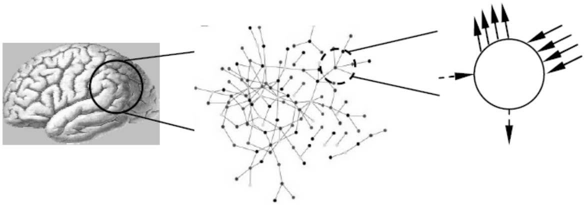


Fig. 1. Distributed computational model of brain

Let G is a simple graph representing a neuro-physiological structure (a brain) and is composed of a set of nodes N having overall conscious output set ω_G. The row vector $<e>$ represents the environmental input (excitation) to G. The set ω_{Gn} represents outputs of $n \in N$ to environment (retaining history), whereas ω_{G^*} denotes the immediately past set of outputs of G to the environment. Let, the set of random environmental sensory input to a conscious brain is given by $I_E = \{X_E\ (<e>, \omega_{G^*})\}$ where, $X_E: (<e>, \omega_{G^*}) \rightarrow R$ represents random environment variable and, the input to a node n from environment is $i_{en} \in I_E$. Let $S \subset \mathbb{Z}$ is a set of excitation states in $G = (N, L_G)$, where $L_G \subset N^2$. Suppose, $I_{\beta n}$ is a set of all inter-node input signals coming to $n \in N$ from other nodes (excluding environment) and, the entire set of input to a node

n is given by, $I_{\alpha n} = (I_{\alpha} \subset I_E) \cup I_{\beta n}$. Furthermore, the inter-nodal signal generated by a node n due to excitation is I_{nt}. Each node in G has a set of output channels given by O_n, a transformation function represented by $\sigma(.)$ and, a transmission function at time t given by, $\gamma_t(.)$. It is evident that set $I_{\alpha n}$ contains every possible input to a node in brain including excitations from the environment. The normal operation of a brain having consciousness is dependent upon two parameters such as, (1) all the nodes in $G = (N, L_G)$ should be operational and, (2) inter-nodal signal propagation should be normal. The state of consciousness can be dynamically varying over time and the dynamic state of consciousness can have two bounds, a maxima (v) and a minima (u), where (u, $v \in \mathbb{Z}$). The boundary state of consciousness $u < 0$ of a node $n \in N$ is considered if the function of node is impaired due to some physical or environmental conditions. It is evident that, a person having normal state of consciousness produces output to environment due to an excitatory input to n by utilizing a subset of nodes $N_e \subset N \setminus n$ together with the node $n \in N$. Hence, the dynamic selections of N_e depending upon the types of environmental excitations and, the inter-nodal information transactions to produce conscious output are two important steps to consider. Let, $\forall n \in N$, $f_n(.)$ is a selection function at n depending upon different types of excitations originated from environment. It dynamically determines the subset of nodes required to produce a conscious output due to a particular input either from environment or the excitations from other nodes in a brain having normal state of consciousness.

The excitation function due to an input to a node of the brain having normal state of consciousness is defined as, $\delta : I_{\alpha n} \rightarrow S$ such that, $\delta.(\delta^{-1}) = 1$. As a result, there will be a local and temporary excitation at node n represented by λ_n and, the excitation is internalized by the respective neuronal node through a fuzzy membership function $\mu_n(.)$. There exists a k such that, $\delta(i_x) = ki_x$ and, $k = 1$ if $i_x \in I_{\beta n}$, otherwise, $0 < k < max(S)$. The output of a node as well as the set of nodes (i.e. brain) due to an excitation is controlled such that, ω_{G^*}, $\omega_{Gn} \subset \omega_G$ and, $\omega_{G^*} \cap \omega_{Gn} \neq \phi$. Thus, for the entire brain represented by G, $\omega_G = \cup_{n\,=\,1,\,|N|}\ \omega_{Gn}$, which is a time-ordered set of output of G. Effectively, the inter-node excitation (information) transactions are controlled by the function $\gamma_t(.)$. Suppose, at time t, $N_e = f_n(.)$. Hence, the dynamics of state of consciousness of a node $n \in N$ can be governed by a triplet function represented as follows ($\forall h \in N_e$, $\lambda_n \in [-u, v]$, $\mu_n(.) \in [0, 1]$ and, $Y \subset P(O_n)$, where $P(.)$ is power set):

$$\boxed{\begin{array}{l} \lambda_n = \sigma(\mu_n(\delta(I_{\alpha n})),\ \omega_{Gn}) \\ f_n : (\delta(I_{\alpha n}),\ \lambda_n) \rightarrow Y \\ \gamma_t : (I_{nt},\ h) \rightarrow \{0, 1\} \end{array}} \tag{1}$$

Let, $(n_a, n_b) \in L_G$ and a channel function in G is defined as, C: $I_{nt} \rightarrow I_{\beta n}$ where, $(i_b \in I_{\beta n}) = C(i_a \in I_{nt})$. The delay-density distribution of the channel is $d_C(x)$ such that, if the propagation delay for a particular signal is τ then, $i_b|_{\tau} = i_a|_{t=0} \int_{0,\tau} d_C(x)\, dx$, where $d_C(x) = d/dx(e^{-x}(cos\ x - sin\ x))$. This indicates that, the natural dynamics of consciousness states at any time t are dependent upon the execution of communication function $\gamma_t(.)$. In addition, the level of inter-node excitation (I_{nt}) generated in node n at time t is equally important for inter-nodal coordination in generating a conscious decision. This is important to note that, it is not necessary to restrict the dynamics of consciousness states by imposing condition as, f_n (.) = f_j (.) for an input from environment.

3.1 Generating Conscious Output

Suppose at time t, a certain excitation has entered into a node $n \in N$ of graph G. Thus, at time $t+a$, (a > 0), the row vector representing a state of consciousness is given by ($m = |f_n(.)|$),

$$\lambda_{\Sigma n}\big|_{t+a} = (\lambda_n, \lambda_1, \lambda_2, \ldots\ldots\ldots \lambda_m) \qquad (2)$$

The output due to excitation is generated at time $t+b$, ($b > a$), from a conscious brain and, it is computed by a function depending upon $\lambda_{\Sigma n}\big|_{t+a}$ as, $\beta_n\big|_{t+b} = g\,(\lambda_{\Sigma n}\big|_{t+a})$ such that,

$$g : \lambda_{\Sigma n} \rightarrow R \qquad (3)$$

It is possible to limit the range of g (.) in a way so that at any time, $\beta_n\big|_t \in [-r, r]$, $r \in \mathcal{Z}+$.

3.2 Evolution of State of Consciousness

The evolution of state of consciousness at time t is highly dependent upon the experiences it has gone through during 0 < t–1 < t. Let, an ordered pair $\psi_{n,t+i} = <I_{\alpha n}\big|_t, \beta_n\big|_{t+i}>$ represents experience in n for i > 0. Thus, the consciousness of a brain with merged experiences can be computed as a finite set, $\omega_G = \{\,\psi_{n,t} : n \in N, t \in \mathcal{Z}+\}$ and, $\omega_{Gn} = \cup_{t\,\in \mathcal{Z}+}\ \psi_{n,t}$. Interestingly, this completely fits into the model of state of consciousness and the respective dynamics.

3.3 Deterministic State of Consciousness

Suppose in a system, all nodes of a brain are functional without physical impairment indicating $\forall t$, $\gamma_t(.) = 1$. The computed value of $\beta_n\big|_t$ for different cases can occur in three ways as follows (considering $|U|= |V|= 1$), (I) $\beta_n\big|_t \in \{-1, 0, 1\}$, (II) $\beta_n\big|_t \in (-1, 0)$ and, (III) $\beta_n\big|_t \in (0, 1)$. An output $\beta_n\big|_t = 1$ indicates that, a deterministic positive conscious decision is made; $\beta_n\big|_t = -1$ indicates that, a deterministic negative conscious decision is made and, $\beta_n\big|_t = 0$ indicates neutrality in conscious decision at time t. The intermediate values of $\beta_n\big|_t$ indicate indeterminism in conscious decision either with positive-bias or with negative-bias depending upon respective signs. It is important to note that, more than one value of $\beta_n\big|_t$ can never occur simultaneously at any single point of time. Hence, if output is following case (I), then it is deterministic state of consciousness at that point of time. On the other hand, the occurrence of case (II) or case (III) at output indicates indeterministic state of consciousness in decision.

3.4 Distributed Global Consensus

The regions of neuronal networks in brain compute and propagate signals from source region to destination region through a series of neuronal firings. The globally consistent conscious output is made by a distributed consensus by different regions in brain in coherence. The information propagation and formation of regions are illustrated in Fig. 2.

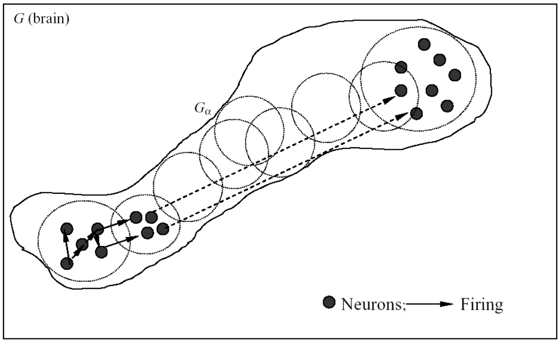


Fig. 2. Schematic diagram illustrating information propagation

Let, $\Delta_j = G_j(N_j, L_{Gj})$ is a sub-graph of G representing a localized region and, for $I_{\alpha n}$, $G_\alpha = \cup_{j = 1, F} \Delta_j$. The regional consensus is defined as a potential given by, $C^{\Delta j} = (1/|N_j|)\int_{\Delta j} d\beta_n$, $n \in N_j$. The distributed global consensus about a conscious output due to an excitation is computed by a distributed consensus algorithm in different regions of brain represented by $D(G_\alpha, C^{\Delta j} : 1 \leq j \leq F)$. An example of simplified definition of generating global consensus about consciousness is, $D(.) = C^{\Delta F} - max(C^{\Delta j} : 1 \leq j \leq F-1)$.

4 Experimental Evaluation

The evaluation is made by computing the state of deterministic consciousness and, the degree of indeterminism in each case. The numerical simulation of the model is implemented considering a set of choice functions governing the dynamics. In nature, the dynamics of deterministic consciousness appear to be stable. In order to reduce rapid excitation and to reduce overshoot/undershoot of the system, a smooth non-linear fuzzy function is chosen. The definition of fuzzy membership function is given as follows,

$$\mu_n(\delta(x)) = \begin{cases} 0.5(1 + \delta(x)^2) \text{ if } x \in I_{\alpha n} \\ 0, \text{otherwise} \end{cases} \tag{4}$$

The characteristic variation of fuzzy function (y-axis) for unit gain (i.e. linear mapping of excitation) for the corresponding varying input vector (x-axis) is illustrated in Fig. 3. Furthermore, the experiments are carried out following two different combinations of inputs. In one case, the input vectors to the system are always deterministic (i.e. {−1, 0, 1}). In another case, the input vectors are made combinatorial in the range [−1, 1]. The input vectors are chosen with uniform randomness in all cases.

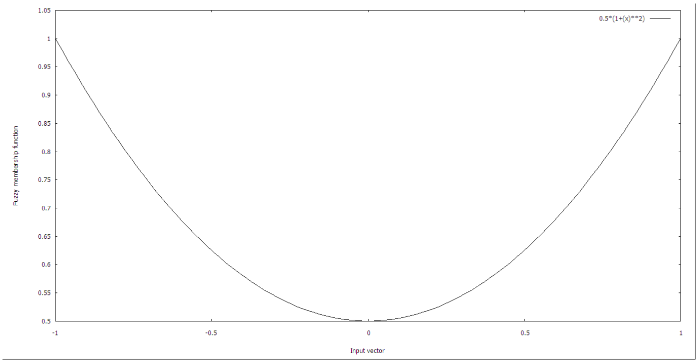


Fig. 3. Characteristics of fuzzy membership function at $k = 1$

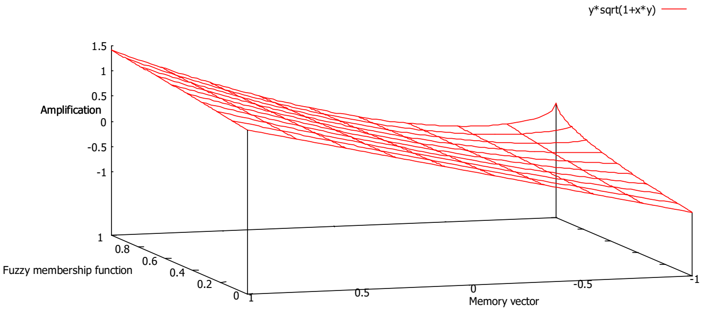


Fig. 4. Characteristic map of constrained $\sigma(x, y)$

According to natural observations, the evolution of state of consciousness is gradual. In order to understand the gradual evolution dynamics of state of deterministic consciousness, a moderate transformation function is chosen. A over amplified transformation function is avoided to eliminate instability in conscious behaviour. The transformation function of a node is chosen as, $\sigma(x, y) = y(1+xy)^{0.5}$, where $x = \mu_n(.)$, $y = avg(\omega_{Gn})$. The transformation is constrained within the limits as, $x \in [0, 1]$ and, the corresponding memory vector $y \in [-1, 1]$. The characteristic surface map of constrained transformation function is illustrated in Fig. 4, where vertical z-axis is $\sigma(.)$, horizontal x-axis is memory and y-axis is $\mu_n(.)$. It is observable from Fig. 4 that, the high and uneven surface of amplification is avoided in the constrained transformation function, which would correlate to the gradual evolution of the deterministic state of consciousness in a system.

4.1 Consciousness without Memory

In this section, a memory-less system is considered, where the state of deterministic consciousness is instantaneous irrespective of dynamics of input vectors. The instantaneous state of deterministic consciousness is computed following continuous averaging method. Thus, in this experiment the parameters are chosen as, $u = v = r = 1$ and, $g(\lambda_{\Sigma n}) = (\lambda_n + \Sigma_{j=1,\mathrm{m}}\lambda_j)\,/\,(m + 1)$. The indeterminism in the consciousness is evaluated by computing relative distance between a state of deterministic consciousness (positive, neutral or negative) and the corresponding computed output. The input vectors are varied into two classes such as, (C1): deterministic input vectors (excitations are in {−1, 0, 1}) and, (C2): combinatorial (or fair) input vectors (where excitations are chosen in [−1, 1] randomly). The comparative study of degree of variations of indeterminism and the corresponding deterministic consciousness are illustrated in Figs. 5-8, where the input vector size is monotonically increased for both classes of input vectors (Figs. 5 and 6 represent size = 2 and, Figs. 7 and 8 represent size = 4).

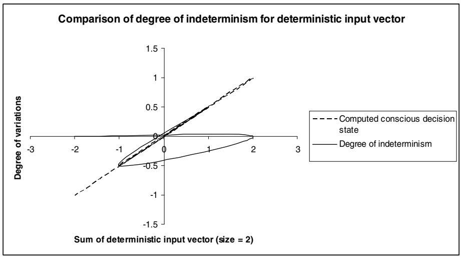


Fig. 5. Variation of consciousness in C1

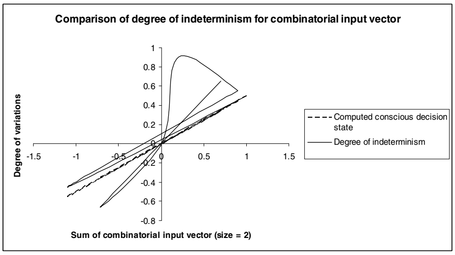


Fig. 6. Variation of consciousness in C2

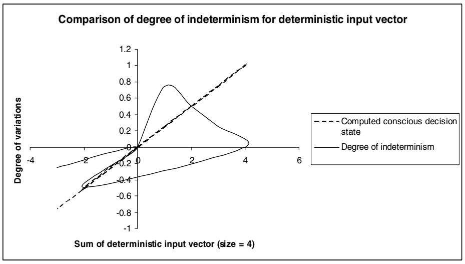


Fig. 7. Variation of consciousness in C1

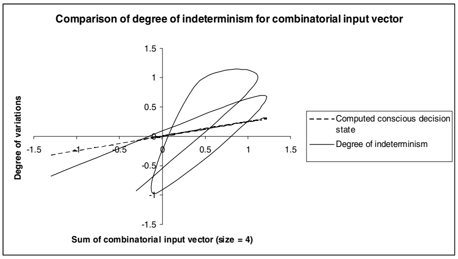


Fig. 8. Variation of consciousness in C2

It is observable that, the conscious decision states are limited within a relatively small bounded domain in comparison to indeterminism. The dynamics of consciousness appears to be closer to linearity in majority cases (not all cases). However, the indeterminism in consciousness is highly non-linear and unpredictable (i.e. chaotic in nature). In addition, the chaotic variations of indeterminism in consciousness tend to increase with the increasing size of combinatorial input vector class (C2). The patterns of variations of deterministic consciousness appear to be unaffected to a high degree for the increasing size of input vectors of both classes.

This is consistent with nature because, the system has zero memory and, the evolution of state of consciousness is stateless.

4.2 Consciousness with Transformation and $\psi_{n,t+i}$

In this section, the experiments are carried out in a system with memory and, the non-linear transformation function is employed to the excitation. In the case of complete memory, the system can store the elements of previous state of consciousness as well as history of input vectors indicating, $y|_t = avg(\beta_n|_{t-1})$ and, $I_{\alpha n} = \Sigma_{j=0,\, t}\, I_{\alpha n}|_j$. The initial values are set as, $y|_0 = 0.5$ and, $k = 1$. The inter-nodal signal transaction values are computed as, $I_{nt} = 0.5(\delta(I_{\alpha n}) + \lambda_n)$ to eliminate over-excitation of transmission. In the case of evaluating a system having memory, the input vector classes are further subdivided into three types such as, (1) fair input vectors (where input values can vary randomly in [−1, 1]), (2) positively-biased input vectors (where input values can vary randomly in [0, 1]) and, (3) negatively-biased input vectors (where input values can vary randomly in [−1, 0]). The initial point of evolution of the state of consciousness of the system is considered to have median value. A relatively low amplification factor of excitation is considered to estimate natural dynamics. The evolution of positive deterministic consciousness in a system having complete memory along with fair input is illustrated in Fig. 9. It is evident that, when input vector size is monotonically increased, the system achieves deterministic consciousness rapidly and the degree of indeterminism reduces to zero. However, if the input vector is made positively-biased, then the state of deterministic consciousness fails to reach unity as illustrated in Fig. 10.

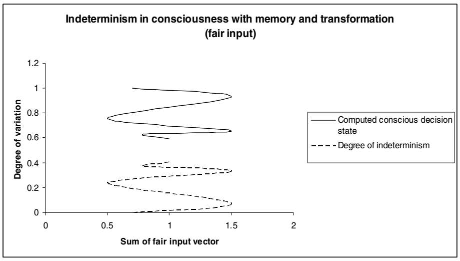


Fig. 9. Consciousness in transformed fair C2

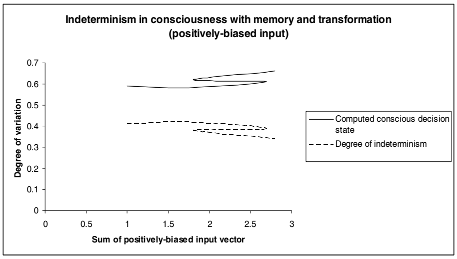


Fig. 10. Consciousness in transformed C2 (+)

The similar effect is observable in Fig. 11, when the input vector is made negatively-biased. In both the cases (with ±biased inputs), a system having complete memory fails to achieve positive deterministic consciousness. Thus, the system having complete memory along with fair inputs can successfully evolve to the positive deterministic consciousness, which is consistent to nature. In other cases, the residual degree of indeterminism remains in the system to varying degrees.

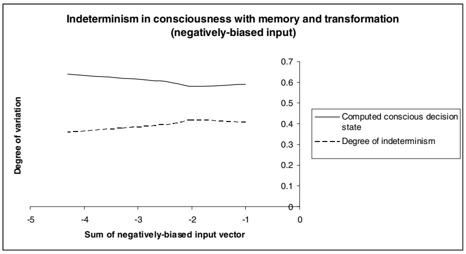


Fig. 11. Consciousness in transformed C2 (–)

5 Conclusion

The state of consciousness is an important aspect of cognitive functions of a neuro-anatomical structure such as, brain. The modeling of dynamics of state of consciousness is central to understanding of cognitive functions. The computational model of state of deterministic consciousness along with its gradual evolution can be constructed considering functional neurophysiology and elements of distributed computing structures. The constructed computational model of deterministic consciousness successfully incorporates scope of memory in determining its evolutionary dynamics. According to the model, the fair environmental excitations and memory are the keys to gradual evolution of consciousness, which is consistent to nature. The close matching of the model to neurological functions can be achieved by appropriate choice functions, where the basic framework of the model remains unchanged.

References

1. Fitch, W.T.: Toward a Computational Framework for Cognitive Biology: Unifying approaches from cognitive neuroscience and comparative cognition. Physics of Life Reviews (2014), doi:10.1016/j.plrev.2014.04.005
2. Reggia, J.A.: The rise of machine consciousness: Studying consciousness with computational models. Neural Networks 44, 112–131 (2013)
3. Fekete, T., Edelman, S.: Towards a computational theory of experience. Consciousness and Cognition 20, 807–827 (2011)
4. Atkinson, A.P., Thomas, M.S.C., Cleeremans, A.: Consciousness: mapping the theoretical landscape. Trends in Cognitive Sciences 4(10) (2000)
5. Bonchek-Dokow, E., Kaminka, G.A.: Towards computational models of intention detection and intention prediction. Cognitive Systems Research 28, 44–79 (2014)
6. Aleksander, I.: Modeling Consciousness in Virtual Computational Machines. Synthesis Philosophica 22(2), 447–454 (2008)
7. Lin, J., Yang, J.G.: Consciousness modeling: A neural computing approach. In: Proceedings of the Third International Conference on Machine Learning and Cybernetics. IEEE, Shanghai (2004)
8. Starzyk, J.A., Prasad, D.K.: A Computational model of machine consciousness. Int. J. Machine Consciousness 3(2) (2011)

9. Arbib, M.A., Caplan, D.: Neurolinguistics must be computational. Behavioral & Brain Sciences 2(3), 449–483 (1979)
10. Poeppel, D., Embick, D.: Defining the relation between linguistics and neuroscience. In: Cutler, A. (ed.) Twenty-First Century Psycholinguistics: Four Cornerstones, pp. 103–120. Lawrence Erlbaum, London (2005)
11. Block, N.: Two neural correlates of consciousness. Trends in Cognitive Sciences 9(2), 46–52 (2005)
12. Mumford, D., Desolneux, A.: Pattern Theory: The stochastic analysis of real-world signals. A K Peters Ltd., CRC Press (2010)
13. Ward, L.: The thalamic dynamic core theory of conscious experience. Consciousness and Cognition 20(2), 464–486 (2011)
14. Mumford, D.: On the computational architecture of the neocortex. II. The role of cortico-cortical loops. Biological Cybernetics 66(3), 241–251 (1992)
15. Rolls, E.T., Deco, G.: Computational Neuroscience of Vision. Oxford University Press, Oxford (2001)
16. Taylor, J.G.: A general framework for the functions of the brain. In: Proceedings of the IEEE-INNS International Joint Conference on Neural Networks, vol. 1, pp. 35–40. IEEE (2000)
17. Tononi, G.: Consciousness as integrated information: A provisional manifesto. The Biological Bulletin 215(3), 216–242 (2008)
18. Breazeal, C., Buchsbaum, D., Gray, J., Gatenby, D., Blumberg, B.: Learning from and about others: Towards using imitation to bootstrap the social understanding of others by robots. Artificial Life 11(1-2), 31–62 (2005)
19. Seth, A.K., Izhikevich, E.M., Reeke, G.N., Edelman, G.M.: Theories and measures of consciousness: An extended framework. Proc. of National Academy of Sciences, USA 103(28), 10799–10804 (2006)
20. Manzotti, R.: Consciousness and existence as a process. Mind and Matter 4(1), 7–43 (2006)

Rotation Invariant Texture Classification Using Principal Direction Estimation

Yulong Qiao and Ying Zhao

School of Information and Communication Engineering,
Harbin Engineering University,
Harbin, 150001, China
wszy198979wszy@163.com

Abstract. The rotation invariant texture classification is an important application of texture analysis. A rotated texture is often perceived by the changed dominant direction. This paper proposes an effective rotation-invariant texture classification method by combining the local patch based method with the orientation estimation. For a texture sample, the Principal component analysis is applied to its local patch to estimate the local orientation, and then the dominant orientation is determined with the maximum value of the local orientation distribution. In order to extract the feature vector, each local patch is rotated along the dominant orientation after circular interpolation. By using the random projection, the local gray value vector of a patch is mapped into a low-dimensional feature vector that is placed in the bag of words model, together with local orientation feature. The simulation experiments demonstrate the proposed method has a comparable performance with the existing methods.

Keywords: Rotation-invariant texture classification, principal component analysis, orientation estimation.

1 Introduction

Texture is a nature attribute of the images from the real world. Texture classification is to classify different texture categories due to our priori knowledge. It's a common issue in computer vision and image processing, and has been widely used in many fields. In the texture classification field, classify textures with rotation invariant is an essential need in many applications.

Conventional method includes polar coordinate transform, Fourier transform, High-order statistics, center-symmetric auto-correlation, local binary pattern[1-2], Gaussian Markov random fields, Radon transform[3-4], Wavelet transform[5], and so forth [6]. In recent years, many researches have been focused on the rotation invariant texture classification (RITC) issue. Among these, patch based method can achieve high accuracy classification rate, while with high computational complexity and no RITC ability. The computational problem is solved in [7], which uses the random projection to map the high-dimensional feature into low dimensional with random measure matrix.

H. Sun et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 329, DOI: 10.1007/978-3-319-12286-1_25

VZ_joint[8] is a patch based method proposed by Varma. To achieve rotation invariant property, the author points out to find a local direction and arrange the intensity values from each circle in local patch. In [9], Zhen uses a structure tensor to estimate the local orientation as a comparison method. The author also proposes a new method ---joint_sort, where the intensity values from local patch are sorted in a descending order. It has a lower accuracy but save the time using to find local direction. However, both of them ignore the global direction. As shown in Fig. 1, as the condition patches with size of 3 pixels, in the patches feature extraction stage, there will be no difference between Fig. 1(a) an Fig. 1(b) in the method of VZ_joint. The same situation also happen between Fig. 1(a) and Fig. 1(c) in the method of joint_sort. Although the overlapping patches can support some orientation information, they don't have enough information of textures with strong global direction.

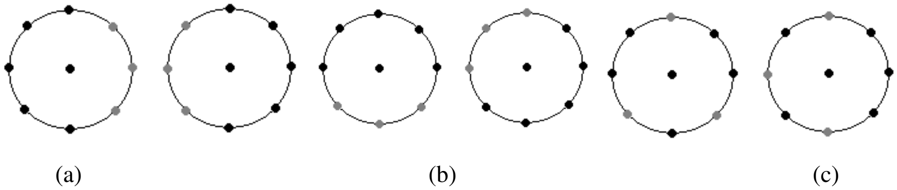


Fig. 1. Three images with size of 3×6 pixels

Textures are rich in orientation information, which can be divided into local orientation and global orientation. As shown in Fig. 2, the white rings point out the local orientation and the black arrows show the mainly global orientation. The local orientation region shows the microscopic structure of texture image, and the global orientation provides the global structure. In order to keep the global and local orientation information at the same time, this paper presents a method combining the local patch and orientation estimate together. The Principal component analysis theory is used to find the local orientation of the micro patch, and the mainly orientation of the texture image is found by the local orientation distribution. Then the feature vectors extracted from each patch are arranged in the mainly orientation, after random projection, the low feature vector and local orientation feature of each patch are used for clustering, training and testing. Experiments have been taken on Outex and CUReT database, which prove the effectiveness of the proposed method.

The rest of this paper is organized as follow: In section 2, we give a brief review of joint_sort and VZ_joint method, and present the method used to estimate the orientation in textures. In section 3, we introduce the proposed method. In section 4, we present the experimental results on benchmark texture datasets. Finally, we draw the conclusions in section 5.

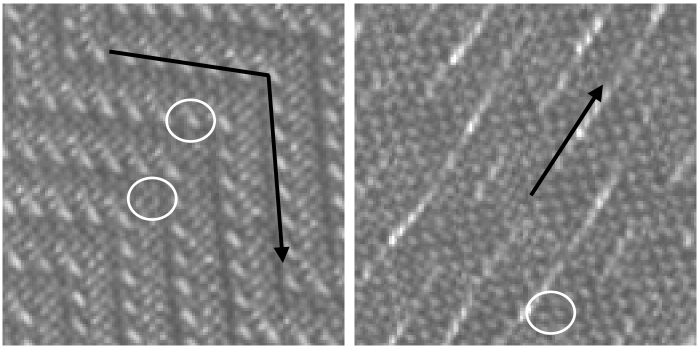


Fig. 2. Two texture images

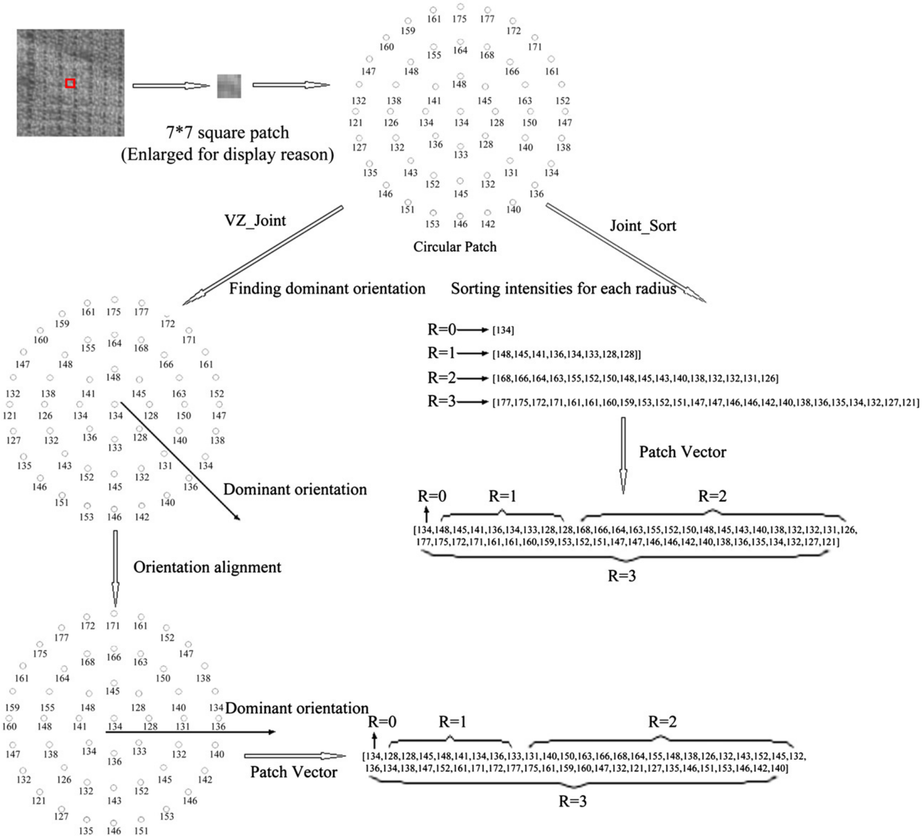


Fig. 3. The difference in feature extraction between VZ_joint and Joint_sort (Source: [4])

2 Backgrounds

2.1 Brief Review of VZ_joint and Joint_sort

VZ_joint is a patch-based method proposed by Varma. The method takes the $N\times N$ square local overlapping patches from texture image to get the N^2 dimensional feature vectors. Then a textons library is built from the train set by clustering and textons histograms are used to classify textures. To address the rotation invariant property, a circle neighborhood is used instead of square in [10], and a local direction achieved by structure tensor [9] is found, then feature vectors are arranged in this direction.

Similar framework is to be found in joint_sort method. The only difference between them is in the feature extraction stage. In joint_sort, the intensity values in each concentric circle are sorted in a descending order, and then concatenate together as a long vector to present the patch. Fig. 3 shows the detail of difference between the two methods.

2.2 The Methods of Orientation Estimate

In [11], an orientation estimation method based on Principal component analysis (PCA) and multiscale pyramid decomposition is proposed. The PCA analysis is applied to find the Maximum Likelihood (ML) estimate of the local orientation [11]. For an image, its gradient image at (x_k,y_k) is denoted as$\nabla f(k)=[\partial f(x_k, y_k) / \partial x, \partial f(x_k, y_k) / \partial y]$, then the gradient image is divided into local blocks (overlapped or non-overlapped). For each $N\times N$ block, we arrange the gradient values by rows to get a $N^2 \times 2$ matrix G, which can be denoted as follows:

$$G=\begin{bmatrix} \nabla f(1) \\ \vdots \\ \nabla f(N^2) \end{bmatrix} \tag{1}$$

Then the Singular Value Decomposition (SVD) is applied on matrix G:

$$G=USV^T \tag{2}$$

And the first column $v1$ of V can be seen as the dominant orientation of the gradient field. Besides, a measure R of accuracy of dominance of the estimate is computed as follows:

$$R=\frac{s_1-s_2}{s_1+s_2} \tag{3}$$

More detail information of the algorithm can be found in [12]. In this paper, to simplify the algorithm, instead of combining the PCA with the multiscale pyramid decomposition, we only use the PCA to estimate the local orientation. From the distribution of local orientation we can get the mainly orientation as the global orientation. And the accuracy measure R is also used as a local feature.

3 The Proposed Patch_Orientation Method

3.1 Patch Based Texture Classification Method Framework

The patch based texture classification method can be mainly divided into two parts: feature extraction and classification. For the training samples, the overlapping patches are extracted, which are used to estimate the local orientation and form the local feature with the mainly orientation of the whole image. After random projections, the textons library is built from low-dimensional feature vectors by k-means clustering. Then the statistical textons distribution histogram of each texture class can represent this class. For example, if there are *C* texture categories with *D* samples per class, and each class has *K* textons, then the textons number of the texton dictionary is *C*×*K*. And the *D* training histograms with the dimension of *C*×*K* will be seen as the representation of this texture class.

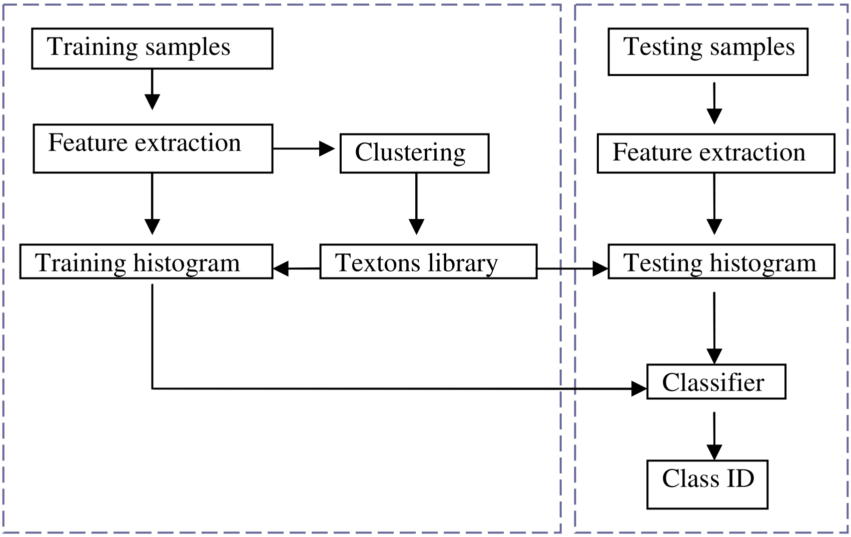


Fig. 4. Patch based texture classification method framework

The classifier employed in this paper is also the *K*-nearest neighbor (KNN). In this theory, according to the similarity measure, the category label of the testing sample is decided by the majority of *k* nearest neighbors in training samples belong. In this paper, the *k* is set to be one, and we use the χ^2 statistics as the similarity measure. Assume H_t and H_r to represent two *N*-dimensional histograms, then the distance *D* between them can be computed like this:

$$D_{\chi^2}(H_t, H_r) = \sum_{i=1}^{N} \frac{[H_t(i) - H_r(i)]^2}{[H_t(i) + H_r(i)]} \tag{4}$$

3.2 Feature Extractions

In this stage, the grey value feature and local orientation feature are extracted in the same time. As shown in Fig. 5, for patch 1, after orientation estimation, the local angle α_1 and estimation measure R_1 are computed from patch 1. Then all local angles from the image generate the orientation histogram, from which we can get the total angle β on the peak of the histogram. And then we interpolate value on the circle region of the patch along the total angle β, and arrange the value circle by circle to get

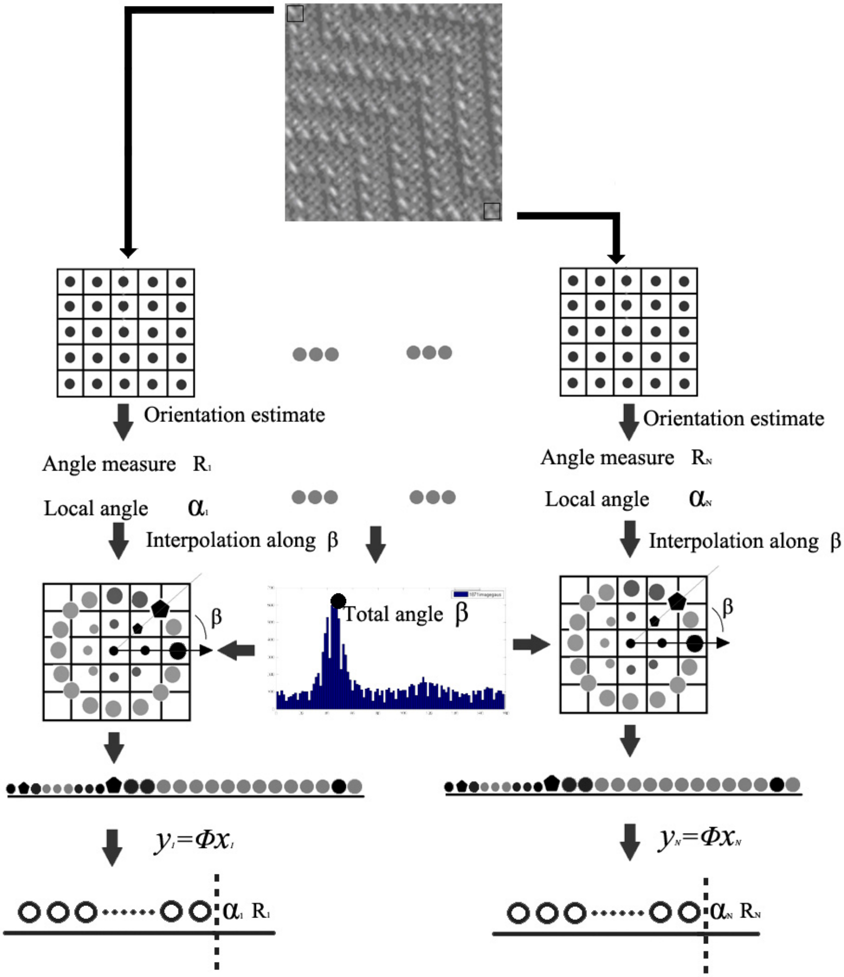


Fig. 5. Feature extractions

the high dimensional vector. To reduce the computational complexity, the random projection is applied to mapping the high dimension patch vector into a low dimension one. The random projection theory is straightforward: For a vector $x = [x_1, x_2,\ldots, x_N]^T$, choose a random matrix $\Phi = [\varphi_{i,j}]_{M\times N}$, the random projection is

$$y = \Phi x \tag{5}$$

Where $N \geq M$, and $y = [y_1, y_2,\ldots, y_M]^T$ is the low-dimensional vector. In [2], inspired by the theory of compressed sensing, the author presents a texture classification method using random projection, where Φ is an independent, zero-mean and unit-variance Gaussian random matrix. As mentioned in [13], there are three kinds of random matrices in compress sensing: Gaussian measurements, Binary measurements, Fourier measurements. According to the experiment results, we choose the matrix Φ to be an independent random matrix with the symmetric Bernoulli distribution $P\left(\phi_{Mi} = \pm 1/\sqrt{M}\right) = 1/2$.

In order to stress the orientation information, after random projection, we combine the low dimensional vector with the local angle α_1 and estimation measure R_1, which are used as the final feature vector. (The local angle α_1 is normalized between 0 and 1.)

4 Experimental Results

To evaluate the effectiveness of the proposed method, we carried out a series of experiments compared with joint_sort and VZ_joint on two benchmark texture datasets: Outex database and Columbia-Utrecht Reflection and Texture (CUReT) database. Each intensity image is individually normalized to have an average intensity of zero and standard deviation. And other experimental setup is the same as in [9].

4.1 Experimental Results on Outex Database

The Outex database contains a large collection of textures with variations to illumination, rotation and spatial resolution. On this section, two test suites are chosen from the sixteen texture classification suites: the Outex_TC_00010 (TC10) test suite and the Outex_TC_00012 (TC12) test suite. Both of them contain the same 24 texture classes at nine angles (00, 05, 10, 15, 30, 45, 60, 75, and 90) with 20 images per rotation angle. In the TC10 test suite, 24×20 images under Inca illuminant with rotation angle 0 are selected as the training samples, and other 24×160 images are selected as testing samples. For TC12 test suite, the training samples chosen are the same as that in TC10 test suite, and other 24×180+480 images under horizon and tl84 illuminant are used for test. Besides, all the training samples in experiments are used to get 40 textons per class.

The experimental results on Outex dataset are listed in table1. The results of VZ_joint and Joint_sort method are copied from [9], and the results of CLBC_CLBP are copied from [14]. It can be seen that the proposed method has a better performance than Joint_sort, and a little lower accuracy than VZ_joint. But the proposed method has a comparative time cost with Joint_sort. If we don't consider the time cost, and use the 360°circular interpolation, the classification accuracy on TC10 of proposed method can achieve 99.193%.

Table 1. Experimental result on Outex database

	Pathc5			Patch7		
	Tc10	Tc12t	Tc12h	Tc10	Tc12t	Tc12h
VZ_joint	-	-	-	98.51	97.45	98.35
CLBC_CLBP	98.83	93.59	94.26	98.96	95.37	94.72
Joint_sort	-	-	-	99.19	94.88	96.82
Proposed method	98.33	92.29	95.67	99.04	97.08	97.50

4.2 Experimental Results on CUReT Database

The Columbia-Utrecht (CUReT) database contains images of 61 materials images under different viewing and illumination conditions. To compared with other methods, the same 92 images per classes are selected and a 200×200 central region is cropped from each images. Besides, all the images are converted to grey scale. In the experiments, *L* samples per class are chosen as the training samples. The first 23 images of each class are used for clustering to get 40 textons per class. And other (92-*L*)×61 images are selected for test.

Table 2. Experimental results on CUReT database

	L			
	46	23	12	6
VZ_joint[4]	97.51±0.75	94.27±1.63	89.00±2.26	80.22±3.93
Joint_sort[4]	96.93±0.95	93.00±1.92	86.28±3.11	76.24±4.16
Proposed	95.43±0.38	90.62±0.56	83.43±0.84	72.88±1.07

Table2 shows the experiment results on CUReT database. To get the significant difference, the experiments are taken over a thousand random splits, and the accuracy represents the average classification rates and standard deviation. It is noted that the proposed method achieves worse results than others. It may be the effect of random projection, as mentioned in [7], the random projection can reduce the computational complexity, but also reduce the accuracy of algorithm at the same time.

5 Conclusion

In the proposed method, the local orientation is combined with the local patch based method. The local orientation estimated by SVD of the gradient image patch is used in two ways, one is for getting the dominant orientation, and the other is gathered into the final feature vector. To reduce the computational complexity, random projection is applied to mapping the high dimension patch vector into a low dimension one. The experiment results show the proposed method has a better performance than joint_sort method on Outex database, and has a little worse result in CUReT database. Although the random projection can reduce the computational complexity, the classification accuracy rate is not satisfactory enough. More experiments should be taken to find whether the random projection affects the performance of proposed method on CUReT database.

Acknowledgments. This work is partially supported by National Natural Science Foundation of China 61371175, Postdoctoral Science-research Developmental Foundation of Heilongjiang Province LBH-Q09128, and Fundamental Research Funds for the Central Universities HEUCFQ1411. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

References

1. Mäenpää, T., Ojala, T., Pietikäinen, M., Soriano, M.: Robust.: Texture Classification by Subsets of Local Binary Patterns. In: Proc. 15th Int'l Conf. Pattern Recognition, vol. 3, pp. 947–950 (2000)
2. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns. IEEE Trans Pattern Analysis and Machine Intelligence 24(7), 971–987 (2002)
3. Arivazhagan, S., Ganesan, L., Subash Kumar, T.G.: Texture classification using ridgelet transform. Pattern Recognition Letters 27(16), 1875–1883 (2006)
4. Pan, W., Bui, T.D., Suen, C.Y.: Rotation invariant texture classification by ridgelet transform and frequency-orientation space decomposition. Signal Processing 88(1), 189–199 (2008)
5. Kaganami, H.G., Ali, S.K., Zou, B.: Optimal approach for texture analysis and classification based on wavelet transform and neural network. Journal of Information Hiding and Multimedia Signal Processing 2, 33–40 (2011)
6. Zhang, J.G., Tan, T.N.: Brief review of invariant texture analysis methods. Pattern Recognition 35(3), 735–747 (2002)
7. Liu, L., Fieguth, P.: Texture classification from random features. IEEE Transactions on Pattern Analysis and Machine Intelligence 34(3), 574–586 (2012)
8. Varma, M., Zisserman, A.: A statistical approach to material classification using image patch exemplars. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(11), 2032–2047 (2009)
9. Li, Q., Zhang, L., You, J., Zhang, D., Liu, W.: Is local dominant orientation necessary for the classification of rotation invariant texture. Neurocomputing 116, 182–191 (2013)

10. Varma, M.: Statistical Approaches to Texture Classication. Oxford, Jesus College (2004)
11. Feng, X., Milanfar, P.: Multiscale Principal Components Analysis for Image Local orientation estimation. In: The 36th Conference on Signals, Systems and Computers, pp. 176–180 (2002)
12. Feng, X.: Analysis and approaches to image local orientation estimation, California Santa Cruz (2003)
13. Candès, E.J.: Compressive sampling. In: International Congress of Mathematicians, ICM, vol. 3, pp. 1433–1452 (2006)
14. Zhao, Y., Huang, D.-S., Jia, W.: Completed Local Binary Count for Rotation Invariant Texture Classification. IEEE Transactions on Image Processing 21(10), 4492–4497 (2012)

An Improved Spatial Histogram and Particle Filter Face Tracking

Dingli Yang[1], Yulin Zhang[1], Rendong Ji[1,2], Yazhou Li[1] ,
Liqun Huangfu[1], and Yudong Yang[1]

[1] Faculty of Electronic and Electrical Engineering, Huaiyin Institute of Technology,
Huai'an 223003, China
[2] College of Science, Nanjing University of Aeronautics and Astronautics,
Nanjing 210016, China
yangdingli@163.com

Abstract. Because uniform division spatial histogram can not finely divide the data in relatively concentrated areas, it can not accurately track human faces. A new face tracking method which combines an improved spatial histogram with particle filter is proposed. In this method, non-uniform division is proposed. Histogram data in relatively concentrated areas can be divided finely, and histogram data in relatively sparse areas can be divided roughly. Simultaneously, a new re-sampling method is proposed in order to solve the "particle degradation" and "particle depletion". If many duplicate particles occur, keep a particle, remove other particles. In order to ensure that the total number of particles is N, particles must be selected randomly in the vicinity of the particles which have a large weight. Experiments show that its tracking performance is very good when target color is similar to the scene color and obstructed partly or completely, or under the complex non-linear, non-Gaussian situations.

Keywords: spatial histogram, non-uniform division, particle filter, re-sample, face tracking.

1 Introduction

Target tracking has been widely applied in video surveillance, motion capture, robot localization etc. Face tracking in image sequence has broad application prospect in video surveillance, face recognition, human computer interaction, and video retrieval etc. At present, there are many face tracking methods [1-4]. But most of the methods are affected by different face pose, changed illumination, complicated background. The color histogram is a common target modeling method. Its advantage is simple calculation, and high performance in real time. It has been applied in target tracking combined with particle filter [5, 6]. Characteristic of color distribution is stable and insensitive to the change of object posture and deformation of the nonrigid objects, and it is invariant to rotation and scale, but it is easy to lose the target when the scene contains the object which's color distribution is similar to the target. To solve this problem, reference [7] depicts system state model with a simple random drift model,

H. Sun et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 329, DOI: 10.1007/978-3-319-12286-1_26

and it establishes systemic observation probability model defined by the similarity of the spatial histogram. Reference [8] proposes a new similarity measure method based on spatial histogram. In this method, spatial distribution of pixel corresponding to each section in spatial histogram is seen as Gaussian distribution, and its mean and covariance matrix are calculated by all the pixel coordinates in this section. After its mean and covariance matrix are obtained, similarity of spatial distribution is calculated by Jensen-Shannon Divergence (JSD), but similarity of color features is calculated with histogram intersection method which has a strong ability to distinguish color features. Reference [7] and reference [8] propose the methods which combine spatial histogram with particle filter. Particle filter algorithm is a non-parametric Monte Carlo simulation method to achieve recursive Bayesian filtering. In these methods, particles, a set of random samples with weights, represent the state of the system posterior probability density function of the system. The estimation of the posterior density function is obtained by these samples. With the increasing of the sample number to the infinity, particle filter can achieve the optimal Bayesian estimation, and it is generally equivalent to the posterior probability density describing by functions.

In these above method, spatial histogram combines color information of the target and the spatial distribution information. It has a stronger ability to identify targets than the traditional color histogram. At present, there are many face tracking methods based on similarity of spatial histogram. But color histogram interval is always divided at equal interval in these references. For example, it is divided at equal interval in reference [1]. In this method, histogram data in relatively concentrated areas can not be divided finely. Thus, the target can not be identified accurately. Simultaneously, since histogram data in relatively sparse areas is divided finely, it increases calculation time. Thus, a new improved method is proposed, which combined non-uniform spatial histogram with particle filter. In this method, histogram data in relatively concentrated areas can be divided finely, and histogram data in relatively sparse areas can be divided roughly. At the same time, a new re-sampling method is utilized. Experiments show that this face tracking method which combines improved spatial histogram with particle filter is more robust, more accurate, and more efficient.

2 Similarity Measure Method of a Non-uniform Spatial Histogram

Color feature is the basic feature of the target tracking. It is stable to partial occlusion, rotation and change in scale. Color distribution is divided into B level in reference [7]. B is the color quantization level and it is quantized uniformly. Quantization interval Δw is equal to $\frac{256}{N}$. $b(l_m): R^2 \to \{1,2,...,B\}$, which is the quantization function, indicates that color values of the pixel which position is l_m is quantized and is assigned into appropriate color level. In order to make a classification finely, Δw will decrease in some sections where probability is greater. It will increase in some sections where probability is smaller. Thus, quantization function $b(l_m): R^2 \to \{1,2,...,B\}$ is improved to $b'(l_m): R^2 \to \{1,2,...,B\}$, which is a new proposed method. When it is quantized, the i-th level pixel value range is [$\sum_{u=1}^{i-1} \Delta w_u$, $\sum_{u=1}^{i} \Delta w_u$]. In this formula, $\Delta w_u = \frac{256}{N}(1 + \frac{1}{N} - p^u)$. If i=1, its range is [0 , Δw_1]. In above formula, the greater probability is, the smaller the range is, and Δw is smaller. Moreover, $\sum_{u=1}^{B} \Delta w_u = 256 + \frac{256}{N} - \frac{256}{N}(p^1 + p^2 + ... + p^u) = 256$. This is division of the chromic interval. Assuming that a target state is the X, $p_l = \{p_l^u\}_{u=1,2,...,B}$ is defined as $p_l^u = C\sum_{m=1}^{M} k(\| \frac{l-l_m}{h} \|)\delta(b'(l_m) - u)$, which represents color distribution. Among the above formula, l is target center which is determined by the target state X. The M represents the total number of pixels in the target area. The h represents the window size of the target area. $k(.)$ represents the kernel function(Gaussian kernel function is selected generally). Constant C is equal to $1 \Big/ \sum_{m=1}^{M} K(\| \frac{l-l_m}{h} \|)$. $b'(l_m)$ is the improved quantization function. Assuming that the i-th sample X_k^i color distribution of the target state X_k is $\{p^u\}_{u=1,2,...,B}$, and color distribution of reference target is $\{q^u\}_{u=1,2,...,B}$, Similarity of candidate target and reference target can be represented using Bhattacharyya coefficient. Bhattacharyya coefficient is defined as $\rho[p^u, q^u] = \sum_{u=1}^{B} \sqrt{p^u q^u}$. Thus, similarity measure function is defined as $D(p,q) = \sqrt{1 - \rho[p^u, q^u]}$. In the above formula, if $D(p,q)$ is smaller, the distance will be smaller.

If the similarity measure functions only utilize color information, tracking will fail when the target is covered by the object which color is similar to the background. Therefore, spatial information of the target color distribution is also utilized. Spatial histogram comprises spatial mean and covariance of each bin in histogram. It can improve the robustness of the tracking because the spatial information can get more information about target.

Second-order spatial histogram of a image can be represented as: $h(b) =< p^b, \mu_b, C_b >, b = (1,2,...,B)$. Among the above formula, p^b is the probability of the pixel falling into the interval b. μ_b and C_b are the mean vector and covariance matrix of the pixel coordinates in the interval b respectively. The similarity measure of the histogram about candidate target and reference target can be represented by the weights of their similarity. It is defined as: $\rho(h,h')=\sum_{b=1}^{B}k\cdot\exp\{-\frac{1}{2}(\mu_b-\mu_b')^T(C_b^{-1}+C_b'^{-1})(\mu_b-\mu_b')\}\sqrt{p^bq^b}$ As can be seen from the above equation, the spatial information is integrated into similarity measure function of spatial histogram. Therefore similarity measure function of spatial histogram is defined more strictly and reliably than traditional similarity measure function based on color histogram.

3 Particle Filter

Particle filter is a method for processing posterior probability density $p(x_t \mid z_t)$ and observed probability density which distribution is non-Gaussian. Its principal method is that the probability distribution is represented by a group of sample set with weight. $\mathrm{X}=\{(\mathrm{x}^{\mathrm{n}},\mathrm{w}^{\mathrm{n}}) \mid \mathrm{n}=1,2,...,\mathrm{N}\}$ is sample set, w is the weight. Assuming x_t is the target state variable at time t, $z_{1:t}$ is the all observed variable at cut-off time t. Then, a set of particles with weight $\mathrm{X}_{\mathrm{t}}=\{(\mathrm{x}_{\mathrm{t}}^{n},\mathrm{w}_{\mathrm{t}}^{\mathrm{n}}) \mid \mathrm{n}=1,2,...,\mathrm{N}\}$, $\sum_{n=1}^{N}\mathrm{w}_{\mathrm{t}}^{\mathrm{n}}=1$ are obtained by sampling from the posterior probability density $p(x_t \mid z_{1:t})$. In order to solve the problem of "particle degradation" and "particle exhaust" in particle filter algorithm, re-sample is done after each tracking [9].

Because the weight of each particle is proportional to its own likelihood probability $p(z_t \mid x_t^n)$ completely, the ultimate state of the target at time t can be estimated using the weighted average of all particles. It is expressed as:

$$E(X_t)=\sum_{n=1}^{N}w_t^nx_t^n \tag{1}$$

The likelihood probability of particle in target tracking is obtained by calculating the similarity distance between the particle and the target model. Assuming that the target model is $h'(u)=\{q_u,\mu_u',C_u'\},u=1,2...B$, the spatial histogram of the n-th particle is defined as $h(u)=\{p_u^n,\mu_u,C_u\},u=1,2...B$. Then the similarity between the particle and the target model is defined as: $\rho(h,h')=\sum_{u=1}^{B}k\bullet\exp\{-\frac{1}{2}(\mu_u-\mu_u')^T(C_u^{-1}+C_u'^{-1})(\mu_u-\mu_u')\}\sqrt{p_u^nq_u}$ Then the Bhattacharyya distance between the particle and the target model is: $d(h,h')=\sqrt{1-\rho(h,h')}$. The likelihood probability of particle is: $p(z \mid x^n)\propto\exp(-d^2/(2\sigma^2))$. where σ is the standard variance of the observed noise.

As can be seen from the above equation, the more similar the particles and the target model are, the smaller the distance between the particles and the target model is, and vice versa. Therefore, likelihood probability of the particle is very high, which is nearby the center of the target, or has little change in the scale, and vice versa. The ultimate state of the target is obtained by these particles with weight. Therefore, the measure method need to be found, which is very strong to distinguish similarity. By this method, the similarity of particle and target model is very high when particle is nearby the center of the target, or has little change in the scale. The similarity of particle and target model is smaller when particle is away from the center of the target, or has big change in the scale. Thus, this method can improve the accuracy and stability of the target tracking.

The method of face tracking based on particle filter mainly includes particles initialization, re-sampling, prediction, observation, target status updating, and so on. Assuming the system state vector is X = (x, y, w, Δx,Δy,Δw), Where x, y indicates the center position of the face region, w indicates the width of the face region, Δx, Δy indicates the displacement change of face movement, Δw indicates the width change of the face region.

(1) Initialization

Firstly, 100 particles are selected randomly near the region of the face. Spatial histogram of every particle is calculated. Then, Bhattacharyya distance between particle and the target of human face is calculated. The weight of each particle is calculated by this formula: $w_t = \frac{1}{\sqrt{2\pi\sigma}}\exp(\frac{-d^2}{2\pi\sigma^2})$. The w_t is normalized by $w_i = \frac{w_i}{\sum_{n=1}^{N} w_n}$.

(2) resampling

After iterating a few times by above formula, the weight of the most of the particles becomes too small. Then, this can cause "particle degradation ". In order to prevent particle degradation caused in dissemination process, particles must be re-sampled. Particle with high weight will generate many new particles instead of the particles which weight is low.

Assuming the particle set is $\{x_t^n, w_t^n, c_t^n, n = 1,2,...N\}$ at time t, where c_t^n is cumulative probability corresponding to the weight of particles. Namely $c_t^n = c_t^{n-1} + w_t^n$. Then the random numbers u_1 distributed uniformly in [0, 1] is generated

```
  for j=1:N
      $\mathrm{u}_\mathrm{j} = \frac{\mathrm{u}_1}{N} + \frac{1}{N}(j-1)$
    while $\mathrm{u}_\mathrm{j} > c_t^n$
        n=n+1;
    end while
  $x_t^{j*} = x_t^n$
end for
```

Although this method can solve the particle degradation, however, particles with high weights are selected many times so that the obtained sample will contain many duplicate particles. This will lead to the loss of particle diversity. This is so-called "particle exhaust". If many duplicate particles appear, only one particle is retained, and other particles are removed. In order to ensure that the total number of particles is N, some particles must be randomly selected in the vicinity of the particles with high weights so that it supplements those particle which is removed.

(3) prediction
The first order autoregressive equation is applicable to the situation that moving target has obvious trend of voluntary movement. The formula is:

$$x_k = Ax_{k-1} + Bv_k \tag{2}$$

Where A, B are a constant matrix, they can be obtained by experiment or experience. x_k is state vector of the moving target at time k. v is normalized noise. Bv represents variance of Gaussian noise, which indicates range of distribution of particle when it is applied to the image. In the above formula, the first part of the right part is sure. The second part is random.

(4) Observation
According to the above predictive method, a new set of particles are obtained after undergoing the dissemination of the particle. Then, observed value and the normalized weight value w_t^n of the each particle are calculated.

(5) Face state updating
The final state of the face can be estimated as:

$$E(x_t^n) = \sum_{n=1}^{N} w_t^n x_t^n \tag{3}$$

4 Face Tracking Method Based on Improved Spatial Histogram and Particle Filter

In this paper, face tracking is realized by improved spatial histogram combined with particle filter algorithm. Specific procedures of this algorithm are as follows:

(1) Particle initialization: At initial time (*k*=0), reference target is selected manually in the initial frame. The color distribution $\{q^u\}_{u=1,2,\ldots,B}$ of the reference target is calculated. Statistical spatial information is obtained. Mean vector and covariance matrix of the pixel position falling into each interval are calculated. Then sample sets of the initial state $\{X_0^i, \frac{1}{N}\}_{i=1}^{N}$ are established according to the prior distribution $p(X0)$.

（2）Particle status transfer: At time $k(k>0)$, According to random drift model and state of particle X_{k-1}^{i}, state of the particle X_{k}^{i} is predicted.

（3）Particle weight calculation: Particle weight is calculated according to $d(h,h')$, and it is normalized:

$$W_{k}^{i}=W_{k}^{i}/\sum_{i=1}^{N}W_{k}^{i} \tag{4}$$

（4）Estimation of the target state: The minimum mean square error estimation of the target state at time k is calculated by this formula: $X_{k}=\sum_{i=1}^{N}W_{k}^{i}X_{k}^{i}$.

（5）Resampling: N samples are re-extracted from the sample set according to the sample weight. Specific steps are as follows:

① Calculating cumulative weight value of the sample collection $\{X_{k}^{i},W_{k}^{i}\}_{i=1}^{N}$: $C_{k}^{i}=C_{k}^{i-1}+W_{k}^{i}$.

② Then random numbers u_{1} uniformly distributed in [0,1] is generated

③ Searching for minimal j from the sample set when $u_{j}<C_{k}^{i}$. And make $X_{k}'^{j}=X_{k}^{i}$.

④ Form the new set of particles $\{X_{k}^{i},W_{k}^{i}\}_{i=1}^{N}=\{X_{k}'^{i},\frac{1}{N}\}_{i=1}^{N}$.

（6） making $k=k+1$, Returning to step (2).

5 Experimental Analysis

In this experiment, CPU is Intel® Core™ i5-3470 CPU @3.20GHz 3.20GHz, Memory is 4.00 GB. Software is MATLAB 7.11.0(R2010b). Video sequences used in experiment are obtained through the video capture system based on TI DM3730 DSP.

5.1 Improved Histogram

To illustrate the advantages of similarity measure of non-uniform division histogram, at first, uniform histogram and non-uniform histogram are compared in first experiment. See Fig.1, 2 and Fig.3. In Fig.1, the interval of the histogram is uniformly divided. Each interval is 256/N=32. But each interval which interval of the histogram is non-uniformly divided is varied in Fig.2 and Fig.3. The width of the interval from N=1 to N=8 is 36.0000, 33.7922, 34.6807, 34.1647, 31.7675, 31.4383, 18.1801, 35.9765 respectively in Fig.2. The experimental data obtained are shown in Table.1. As can be seen from Table 1, where the probability is larger, the interval of the histogram must be decreased with the method of non-uniform division histogram. Such as $p^7 = 0.5569$, its width of the interval is 32. After the histogram is non-uniformly divided, its width of the interval is 18.18, but $p^7 = 0.0664$. However, most of the data are concentrated in the sixth histogram area. Then $p^6 = 0.5185$. Due to $p^6 > 0.5$, the width of the interval is Re-divided with above method until the probability of each interval is less than 0.5. In Table.1, after the width of the interval is divided secondly, $p^7 = 0.4708$, which is less than 0.5, and its width of the interval is 33.87, it is finished. Then non-uniform division histogram is obtained.

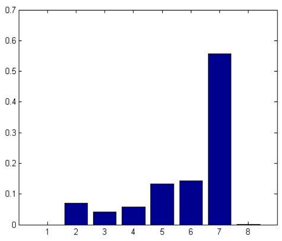


Fig. 1. Uniform division histogram

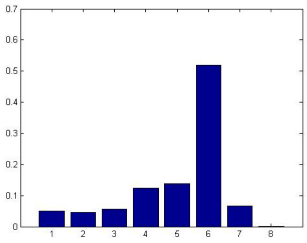


Fig. 2. Non-uniform division histogram firstly

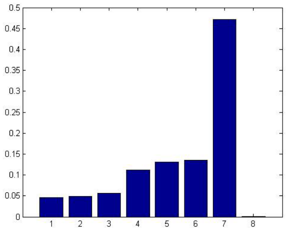


Fig. 3. Non-uniform division histogram secondly

5.2 Tracking Results

The model of reference target expected to track is manually selected in frame 110-Th. The gray of color distribution of the target is Y. Chroma are Cb and Cr. They are divided into eight levels respectively. N is the number of particles. N =100. Face tracking is executed in the same video according to the method of uniform division spatial histogram and non-uniform division spatial histogram respectively. The number of video image sequence is 389. Image resolution is 640 * 360. Environment easily causes interference to the target. Color of obstructions is close to the target color. Performance of two above methods is very good in the previous frame that it is obstructed. In frame 201-Th. Candidate face is obstructed by obstructions completely. In two methods, they can all update the template, and continue to track object. After frame 214-Th, face can be tracked effectively in both methods. However, the error of the both methods is different. See Fig.4. Since non-uniform division spatial histogram is utilized, the ability to distinguish the histogram is stronger, and its error is smaller. In this paper, an error distance between the center position of the tracking target and the center position of a real target in each frame is $e = \sqrt{(x - x_0)^2 + (y - y_0)^2}$.

Table 1. Comparison of histogram of uniform division and non-uniform division

Method		1	2	3	4	5	6	7	8
uniform division	terminal point	31	63	95	127	159	191	223	255
	Width	32	32	32	32	32	32	32	32
	Probability	0	0.069	0.041	0.057	0.132	0.143	0.557	0.001
non-uniform division firstly	terminal point	35	68.79	103.5	137.6	169.4	200.8	219.0	255
	Width	36	33.79	34.68	34.16	31.77	31.44	18.18	35.98
	Probability	0.051	0.046	0.057	0.123	0.139	0.516	0.066	0
non-uniform division secondly	terminal point	33.388	67.93	102.1	134.2	165.7	185.1	219.0	255
	Width	34.384	34.54	34.17	32.06	31.56	19.41	33.87	36.00
	Probability	0.0463	0.049	0.056	0.112	0.131	0.135	0.471	0

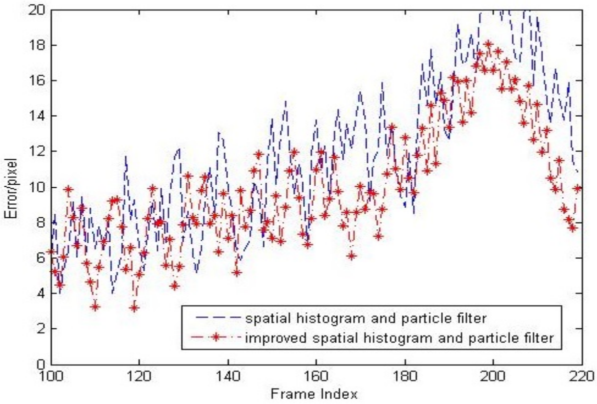
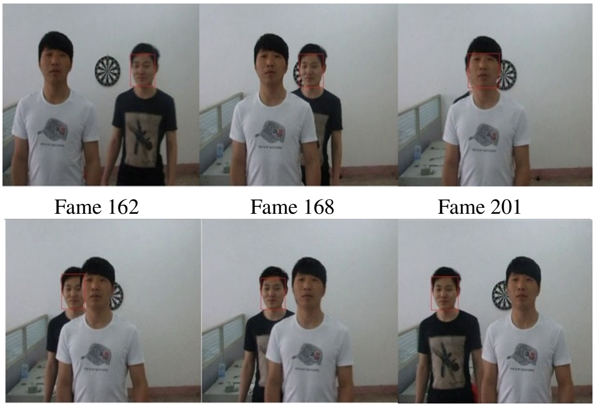


Fig. 4. Comparison of the tracking error



Fame 162 Fame 168 Fame 201

Fame 214 Fame 227 Fame 240

Fig. 5. Results of particle filter tracking with improved spatial histogram

In Fig.5, target can be tracked accurately. It is attributed to the algorithm proposed in this paper, which consider not only the color characteristics of the target, but also spatial information of the characteristics. Its similarity measurement function has a higher ability to identify targets than Bhattacharyya coefficient of the traditional color histogram. And it combines with the merits of particle filter. So it can track the target robustly.

In these above method, spatial histogram combines color information of the target and the spatial distribution information

6 Conclusions

Improved spatial histogram includes not only color information of the target, but also mean vector and covariance matrix of the pixel position of the each histogram interval. It can obtain more characterization than traditional histogram. It utilizes non-uniform division, and where data are concentrated in the color histogram has finer division. Where data are sparse in the color histogram has rough division. It improve the tracking performance. Simultaneously, "particle degradation" and "particle depletion" problem are solved by new re-sampling methods. Finally, the experiments show that it can track the target robustly, and its tracking performance is very good when target color is similar to the scene color and obstructed partly or completely, or under the complex non-linear, non-Gaussian situations.

Acknowledgment. This research was supported in part by a grant from the national natural science foundation of China(GrantNo. 61203056), Foundation of Huaian industrial projects (GrantNo. HAG2013064), Foundation of Huaiyin Institute of Technology (GrantNo. HGB1202).

References

[1] Gao, J., Wang, Y., Yang, H., Wu, Z.: Particle Filter Face Tracking Using Color And Shape Histogram As Clues. Jouranl of Image and Graphics 12(3), 466–473 (2007)

[2] Yang, X., Zhu, H., Deng, Y., et al.: Human Face Tracking Algorithm Based on Particle Filter. Computer Engineering and Applications 44(23), 209–211 (2008)

[3] Yao, H., Zhu, F., Chen, H.: Face Tracking Based on Adaptive PSO Particle Filter. Geomatics and Information Science of Wuhan University 37(4), 492–495 (2012)

[4] Yao, Z., Liu, J., Lai, Z., Liu, W.: An Improved Jensen-Shannon Divergence Based Spatiogram Similarity Measure for Object Tracking. Acta Automatica Sinica 37(12), 1464–1473 (2011)

[5] Wang, Y., Wang, D.: Particle Filter Algorithm for Multi-target Tracking Based on Spatial Histogram. Opto-Electronic Engineering 37(1), 65–75 (2010)

[6] Wang, J., Jiang, Y., Tang, C.: Face Tracking Based on Particle Filter Using Color Histogram and Contour Distributions. Opto-Electonic Engineering 39(10), 32–39 (2012)

[7] Zhang, N., Cai, N., Zhang, H.: Target Tracking Using Particle Filters Based on Spatiograms. Computer Engineering and Applications 47(21), 210–213 (2011)

[8] Yao, Z.: A New Spatiogram Similarity Measure Method and Its Application to Object Tracking. Journal of Electonics & Information Technology 35(7), 1644–1649 (2013)

[9] Zhou, Q.: Study on Face Tracking Algorithm Based on Improved Particle Filtering. Chongqing University, Chongqing (2010)

Joint Encoding of Multi-scale LBP for Infrared Face Recognition

Zhihua Xie and Zhengzi Wang

Key Lab of Optic-Electronic and Communication, Jiangxi Sciences and Technology Normal University, Nanchang, Jiangxi, China
xie_zhihua68@aliyun.com

Abstract. Due to low resolutions of infrared face image, the local feature extraction is more appreciated for infrared face feature extraction. In the current LBP (local binary pattern) feature extraction on infrared face recognition, single scale is encoded, which consider limited local discriminative information. A new infrared face recognition method based on joint encoding of multi-scale LBP (JEMLBP) is proposed in this paper. To consider correlation in different micro-structures, co-occurrence matrix of multi-scale LBP codes is used to represent the infrared face. The experimental results show the recognition rates of infrared face recognition method based on JEMLBP can reach 91.2% under variable ambient temperatures, outperforms that of the classic method based on single scale LBP histogram.

Keywords: Local Binary Pattern, infrared face recognition, multi-scale, joint encoding, co-occurrence matrix.

1 Introduction

Compared with the traditional gray and color face imaging, infrared imaging can acquire the intrinsic temperature information of the skin, which is robust to the impacts of illumination conditions and disguises [1, 2]. Therefore, infrared face recognition is an active research area of face automatic recognition. However, the challenges of infrared face recognition mainly come from the external environment temperature, low resolution and other factors [3].

This paper focuses on robust infrared face recognition under variable environmental temperatures. Many feature extraction methods are proposed for infrared face recognition [4]. Most of the developed approaches make use of appearance-based methods, such as PCA (Principal Component Analysis), LDA (Linear Discriminant Analysis) and ICA (Independent Component Analysis), which project face images into a subspace where the recognition is carried out [6]. Other reported infrared face recognition approaches are based on the use of local-matching: Local Binary Pattern (LBP) [4] and Gabor Jet Descriptors (GJD) [3]. Due to low resolutions of infrared face image, the local feature extraction is more appreciated for infrared face feature extraction, which can be used to get more local discrimination information. LBP histogram representation

H. Sun et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 329, DOI: 10.1007/978-3-319-12286-1_27

was applied to infrared face recognition by Li et al [4], which get a better performance than statistical methods such as PCA and LDA. LBP-based facial image analysis has been one of the most popular and successful applications in recent years [5]. LBP encodes the relative intensity magnitude between each pixel and its neighboring pixels, which can describe the micro-patterns of the image such as flat areas, spots, lines and edges [4, 13]. The main advantage of this approach lies in the robustness of LBP to monotonic photometric changes and in its computational simplicity [5]. Since the impact of external environment temperature on infrared face image is almost a monotonic transform, the LBP can extract robust features for infrared face recognition under different environment situations. In 2011, the method based on single scale local binary pattern was applied to infrared face recognition by Xie et al [7], which gets a better performance than statistical methods such as PCA and LDA.

Recently, multi-scale strategy was introduced into texture representation to depict richer local structure information in different resolutions [8, 9]. Firstly, single-scale LBP histogram features are extracted in each scale separately. Then, the histograms in each scale are concatenated into a final representation. Since the multi-scale strategy always achieves much better performance than single scale. It is interesting to introduce multi-scale strategy for robust infrared face recognition. However, the classical multi-scale strategy each scale is encoded into histograms individually. In fact, the correlation of the micro-structures in different scales consists of discriminative information. To represent the joint distribution of LBP patterns in different scales, we propose infrared face recognition based on the joint encoding of multi-scale LBP patterns (JEMLBP). Compared to single scale histogram, the multi-scale joint co-occurrence strategy can describe stronger local structures in infrared face images under variable environment temperatures.

2 Local Binary Patterns Representation

Local binary patterns were introduced by Ojala [9] which has a low computational complexity and a low sensitivity to monotonic photometric changes. It has been widely used in biometrics such as face recognition, face detection, facial expression recognition, gender classification, iris recognition and infrared face recognition. In its simplest form, an LBP description of a pixel is created by threshold the values of the 3×3 neighborhood of the pixel against the central pixel and interpreting the result as a binary number. The parameters of the original LBP operator with a radius of 1 pixel and 8 sampling points are demonstrated in figure 1. LBP code for center point g_c can be defined as:

$$LBP_{P,R}(x_c, y_c) = \sum_{i=0}^{P-1} 2^i \cdot S(g_i - g_c) \tag{1}$$

$$S(g_i - g_c) = \begin{cases} 1, g_i - g_c \geq 0 \\ 0, g_i - g_c < 0 \end{cases} \tag{2}$$

Where (x_c, y_c) is the coordinate of the central pixel, g_c is the gray value of the central pixel, g_i is the value of its neighbors, P is the total number of sampling points and R is the radius of the neighborhood. As shown in Figure 1, the LBP patterns with different radiuses characterize different size local structures. In other words, a radius R stands for a scale. The parameters (P, R) can be (8,1), (8, 2),(16, 2) and (8, 3) etc.

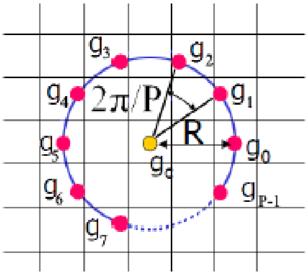
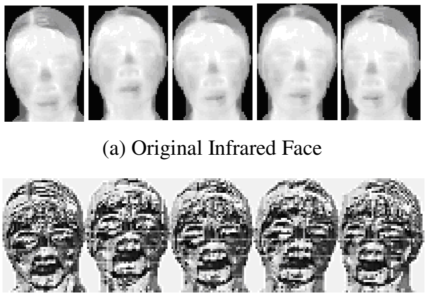


Fig. 1. The parameters P and R of LBP

For the number of sampling points P, The dimension of histogram from LBP codes (LBPH) is 2^P . To reduce the dimension of the LBPH, "uniform patterns" is reserved for representation. An LBP code is defined as uniform if it contains at most two 0-1 or 1-0 transitions when viewed as a circular bit string. If P equals 8, only 59 of the 256 possible 8 bit LBPH bins come from uniform patterns histogram.



(a) Original Infrared Face

(b) LBP Representation

Fig. 2. LBP Patterns Image

Suppose the face image is of size ($M \times N$). After identifying the LBP code of each pixel (x_c, y_c) , by computing the LBP patterns histogram, traditional single scale infrared face recognition methods achieve the final features.

$$H(r) = \sum_{x_c=2}^{N-1} \sum_{y_c=2}^{M-1} f(LBP_{P,R}(x_c, y_c), r) \tag{3}$$

$$f(LBP_{P,R}(x_c,y_c),r)=\begin{cases}1, LBP_{P,R}(x_c,y_c)=r\\ 0, otherwise\end{cases} \tag{4}$$

Where the value r ranges from 0 to 2^P-1.

The center-symmetric local binary patterns feature (CS-LBP) [10, 15], which is a modified version of the LBP texture feature, inherits the desirable properties of both texture features and gradient based features. In addition, they are computationally cheaper and easier to implement. The basic encoding principle of CS-LBP is shown in figure 3. In this paper, we use the CS-LBP to extract micro-structure features in infrared face recognition. In this paper, let *P* equal 8, the dimension number of CS-LBP is 16.

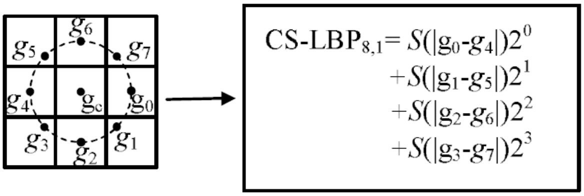


Fig. 3. Basic principle of CS-LBP

3 Joint Encoding of Multi-scale LBP Representation

As shown in figure 1, each scale LBP depicts a local micro-structure in its small scale region. Therefore, single scale LBP can not represent rich local information with large described region. To extract the structure information in different resolutions, multi-scale strategy is usually used in texture classification task [8, 9]. Multi-scale LBP operators are shown in figure 4.

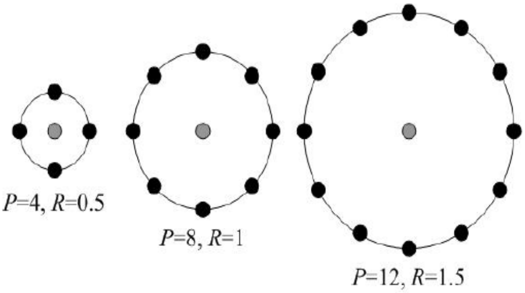


Fig. 4. Multi-scale LBP. R: the radius of sampling circle. P: the total number of sampling points.

Gray-level Co-occurrence Matrix [11] is firstly proposed to calculate co-occurrence of pixel values, but it is sensitive to monotonic photometric changes (external environment temperature impact for infrared face images). Fortunately, the external environment temperature change of doesn't change the relative intensity magnitude in multi-scale LBP. However, the discriminative power of multi-scale LBP is limited by their simply histograms concatenation [12]. In order to represent the joint distribution of LBP patterns in different scales, we propose the joint encoding of multi-scale LBP patterns (JEMLBP) by means of co-occurrence matrix. Considering two scales s1= (P1, R1) and s2= (P2, R2), CMSLBP (s1, s2) can be defined as follow:

$$JEMLBP(s1, s2) = Co(LBP(s1), LBP(s2)) \tag{5}$$

Where $LBP(s1)$ and $LBP(s2)$ are the LBP patterns on scales s1 and scales s2 individually, $Co(LBP(s1), LBP(s2))$ means the co-occurrence statistical operator. A visual illustration of JEMLBP has been shown in figure 5. In this paper, CS-LBP instead of LBP is operated to extract the single scale local patterns. So the dimension number of JMELBP is 256.

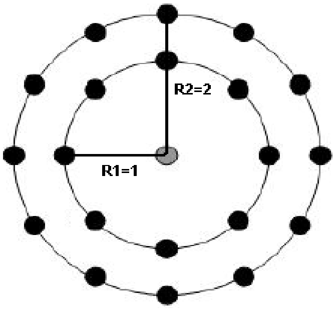


Fig. 5. Spatial co-occurrence matrix of local binary patterns

4 Infrared Face Recognition Method

In this section, the detail realization of our infrared face recognition is introduced. The main steps in our method are listed as follow:

Stage one: Infrared face detection and normalization [1]. After normalization, the resolution of infrared face images is the same.

Stage two: The multi-scale CS-LBP with is applied on the normalized infrared face image to get multi-scale LBP patterns.

Stage three: JEMLBP proposed in Section 3 is applied to build the final representation by partitioning model.

Last stage: The nearest neighborhood classifier based dissimilarity of final features between training datasets and test face is employed to perform the classification task.

In this paper, we use the traditional metric based on chi-square statistic [13, 14]. The dissimilarity of two histograms $(H1, H2)$ can be gotten by:

$$Sim(H1, H2) = \sum_{bin=1}^{n} \frac{\left(H1(bin) - H2(bin)\right)^2}{H1(bin) + H2(bin)} \tag{6}$$

Where n is the dimension of co-occurrence histogram representation of multi-scale LBP.

5 Experimental Results

The infrared data in this paper were captured by an infrared camera Thermo Vision A40 supplied by FLIR Systems Inc [1, 7]. The training database comprises 500 thermal images of 50 individuals which were carefully collected under the similar conditions in November 17, 2006: ambient temperature under air conditioned control with temperature around 25.6~26.3℃. The test data comprises 165 thermal images of one individual which were collected under ambient temperatures from 24.3 to 28.4 °C. The original size for each image is 240×320. After preprocess and normalization, it becomes 80×60. Our experiments take Euclidean distance for final classifier.

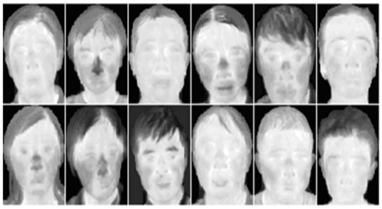


Fig. 6. Part of infrared face database

In our experiments, CS-LBP operator is implemented, parameters of two scales are s1 (8, 1) and s2 (8, 2). The dimension number of histogram extracted by the CMSLBP is 256. To make full use of the space location information, the partitioning is applied to get final features. Five modes of partitioning are used: 1 is non-partitioning, 2 is 2×2, 3 is 4×2, 4 is 2×4, and 5 is 4×4. The recognition results by using LBP, Multi-scale LBP and JEMLBP features with different partitioning modes are demonstrated in figure 7.

It can be seen from figure 7 that for the recognition performance of the method based on JEMLBP outperforms that of the method based on traditional multi-scale LBP and traditional LBP.

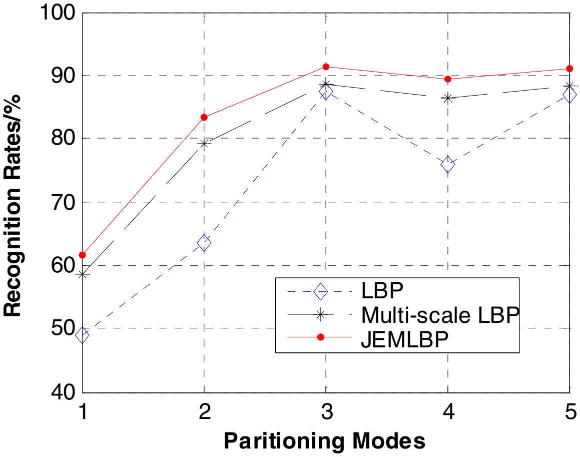


Fig. 7. Recognition results with different partitioning modes

To verify the effectiveness of the proposed features extraction method for infrared face recognition, the three existing features extraction algorithms used for comparisons include traditional LBP histogram [7], traditional multi-scale LBP histogram[8] and PCA+LDA [6].

Table 1. The best Recognition Rate of Different Methods

Methods	Recognition Rates
JEMLBP	**91.2%**
LBP histogram	87.4%
Multi-scale LBP histogram	88.6%
PCA+LDA	33.6%

The best recognition results are demonstrated in Table 1. It can be seen from the Table 1 that co-occurrence matrix of multi-scale LBP can improve the recognition performance of traditional LBP histogram. It is also revealed from Table 1, compared with traditional multi-scale LBP histogram, our proposed JEMLBP can extract more relative information in multi-scale LBP patterns, which contributes to better recognition performance. Therefore, the multi-scale LBP joint encoding is a simple and effective feature extraction method for infrared face recognition.

6 Conclusions

The conventional LBP-based feature as represented by the LBP histogram still has room for performance improvements. Correlation among different scales around the center point could enrich the descriptive power and boost the discriminative power of the descriptors. In this paper, a simple and effective feature is proposed for infrared face recognition. The proposed feature is based on considering joint encoding of multi-scale LBP patterns. Co-occurrence matrix of two scales LBP patterns could provide much more information than their simple concatenation. Our experiments illustrate that JEMLBP method is effective in extraction the discrimination information and the performace of the proposed infrared face recognition method outperforms the muti-scale LBP histogram and traditonal LBP histogram.

Acknowledgment. This paper is supported by Natural Science Foundation of Jiangxi Province of China (No. 20132BAB201052, 20142BAB207029), the Science & Technology Project of Education Bureau of Jiangxi Province (No.GJJ14581) and Nature Science Project of Jiangxi Science and Technology University (2013QNBJRC005, 2013ZDPYJD04).

References

1. Wu, S.Q., Li, W.S., Xie, S.L.: Skin heat transfer model of facial thermograms and its application in face recognition. Pattern Recognition 41(8), 2718–2729 (2008)
2. Li, J., Yu, D.W., Kuang, G.: The Research on Face Recognition Approaches of Infrared Imagery. Journal of National University of Defense Technology 28(2), 73–76 (2006)
3. Hermosilla, G.: A Comparative Study of Thermal Face Recognition Methods in Unconstrained Environments. Pattern Recognition 45(7), 2445–2459 (2012)
4. Li, S.Z., Chu, R., Sheng, C.L.: Illumination Invariant Face Recognition Using Near-Infrared Images. IEEE Trans. on Pattern Analysis and Machine Intelligence 29(12), 627–639 (2007)
5. Ahonen, T., Hadid, A., Pietikäinen, M.: Face description with local binary patterns: application to face recognition. IEEE Trans Pattern Analysis and Machine Intelligence 18(12), 2037–2041 (2006)
6. Hua, S.G., Zhou, Y., Liu, T.: PCA+LDA Based Thermal Infrared Imaging Face Recognition. Pattern Recognition and Artificial Intelligence 21(2), 160–164 (2008)
7. Xie, Z., Zeng, J., Liu, G., et al.: A novel infrared face recognition based on local binary pattern. In: Proceedings of 2011 International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR), pp. 55–59 (2011)
8. Li, W., Fritz, M.: Recognizing materials from virtual examples. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part IV. LNCS, vol. 7575,
pp. 345–358. Springer, Heidelberg (2012)
9. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans Pattern Analysis and Machine Intelligence 7(7), 971–987 (2002)
10. Heikkila, M., Schmid, C.: Description of interest regions with local binary patterns. Pattern Recogntion 42(3), 425–436 (2009)
11. Nosaka, R., Ohkawa, Y., Fukui, K.: Feature Extraction Based on Co-occurrence of Adjacent Local Binary Patterns. In: Ho, Y.-S. (ed.) PSIVT 2011, Part II. LNCS, vol. 7088, pp. 82–91. Springer, Heidelberg (2011)
12. Qi, X., Lu, Y., Chen, S.: Spatial co-occurrence of local intensity order for face recognition. In: Proceedings of 2013 IEEE International Conference on Multimedia and Expo Workshops (ICME), pp. 1–6 (2013)
13. Liao, S., Chung, A.C.S.: Face Recognition with Salient Local Gradient Orientation Binary Pattern. In: Proceedings of 2009 International Conference on Image Processing (ICIP 2009), pp. 3317–3320 (2009)
14. Jabid, T., Kabir, M.H., Chae, O.: Gender Classification using Local Directional Pattern (LDP). In: Proceedings of 2010 International Conference on Pattern Recognition (ICPR 2010), pp. 2162–2164 (2010)
15. Zheng, Y., Shen, C., Hartley, R., Huang, X.: Pyramid Center-symmetric Local Binary/trinary Patterns for Effective Pedestrian Detection. In: Kimmel, R., Klette, R., Sugimoto, A. (eds.) ACCV 2010, Part IV. LNCS, vol. 6495, pp. 281–292. Springer, Heidelberg (2011)

Driving Behavior Analysis of Multiple Information Fusion Based on AdaBoost

Shi-Huang Chen[2], Jeng-Shyang Pan[1], Kaixuan Lu[1], and Huarong Xu[3]

[1] Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, China
[2] Department of Computer Science and Information Engineering, Shu-Te University, Kaohsiung County, 824, Taiwan
[3] Xiamen University of Technology, Xiamen, China
shchen@stu.edu.tw, jengshyangpan@gmail.com,
lukaixuan203@sina.com, hrxu@xmut.edu.cn

Abstract. With the increase in the number of private cars as well as the non-professional drivers, the current traffic environment is in urgent need of driving assist equipment to timely reminder and to rectify the incorrect driving behavior. In order to meet this requirement, this paper proposes an innovative algorithm of driving behavior analysis based on AdaBoost with a variety of driving operation and traffic information. The proposed driving behavior analysis algorithm will mainly monitor driver's driving operation behavior, including steering wheel angle, brake force, and throttle position. To increase the accuracy of driving behavior analysis, the proposed algorithm also takes road conditions into account. The proposed will make use of AdaBoost to create a driving behavior classification model in various different road conditions, and then could determine whether the current driving behavior belongs to safe driving. Experimental results show the correctness of the proposed driving behavior analysis algorithm can achieve average 80% accuracy in various driving simulations. The proposed algorithm has the potential of applying to real-world driver assistance system.

Keywords: Driving behavior analysis, driver assistance system, AdaBoost.

1 Introduction

Due to the continuously increases of the global car ownership, it is unavoidable to raise the traffic density and number of non-professional drivers. This also leads to frequent traffic accidents which have become the first hazard of modern society [1]. Among these traffic accidents, the improper driving behavior habits are an important cause of crashes. Therefore the study of driving behavior analysis has become extremely useful.

Thanks to the modern vehicle manufacturing technology, the un-safety factors of the vehicle itself caused traffic accidents are smaller and smaller proportion. However, the driver personal factors have become the main reason for causing a

H. Sun et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 329, DOI: 10.1007/978-3-319-12286-1_28

traffic accident [2]. Toyota's Masahiro Miyaji, Jiangsu University's Liang Jun [3][4], and other scholars have counted and analyzed the specific reasons of traffic accidents. Fig. 1 shows their research results. From Fig. 1, it is observable that the individual factor account for the accidents are more than 90%. In addition to the driver and pedestrian essential training and education, it also needs to monitor and predict driving behavior of the drivers to prevent the traffic accidents and increase the traffic safety.

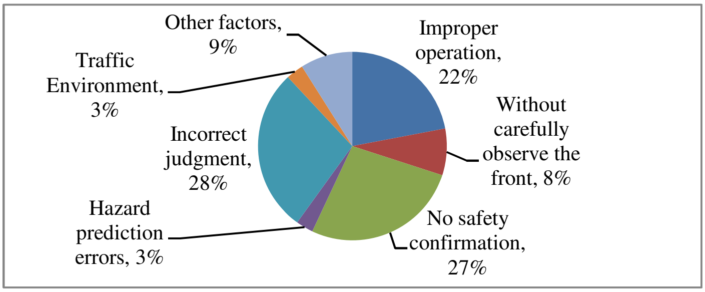


Fig. 1. Distribution of Traffic Accidents Statistics

To meet the requirement mentioned above, it is demanded to analyze driving behavior based on the driver operation and environment. Then it could assess whether the vehicle is in a safe state or not. If necessary, the vehicle system should make the appropriate tips to ensure that the car in a safe condition. Therefore, there are important theoretical significance and application value on the research of driving behavior analysis based on driving operation information and the traffic situation. This paper will try to use the AdaBoost algorithm to complete driving behavior analysis.

The remaining sections are organized as follows. Section 2 brief introduces the AdaBoost theorem. The detail of the proposed driving behavior analysis algorithm is presented in Section 3. Section 4 shows experimental results and analysis. Finally, Section 5 concludes this paper.

2 AdaBoost Basic Principle

AdaBoost is a classical classification machine learning algorithm. The basic principle of AdaBoost algorithm is to use a large number of weak classifiers combined together by a certain method, form a strong classifier which has a strong ability of classification [5][6][7]. Strong classifier generated as follows:

Assuming given a two-classification training data set:

$$T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\} \tag{1}$$

where each sample composed by the instance and flags. Instance $x_i \in \mathrm{X} \in R^n$, flags $y_i \in Y = \{-1,+1\}$, X is the instance space, Y is flags set. AdaBoost use the follow algorithm to generate the strong classifier.

Algorithm .

Input: training data set $T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$, where $x_i \in \mathrm{X} \in R^n$, $y_i \in Y = \{-1,+1\}$; weak learn algorithm;

Output: Strong classifier *G*(*x*).

(1) Initialization of the weight value distribution of the training data

$$D_1 = (w_{11}, \cdots, w_{1i}, \cdots, w_{1N}),\ w_{1i} = \frac{1}{N},\ i = 1, 2, \cdots, N \tag{2}$$

(2) $m = 1, 2, \cdots, M$ (m is the times of train.)

(a) Using training data set has the weight distribution D_m to learn, get the basic classification

$$G_m(x) : X \rightarrow \{-1,+1\} \tag{3}$$

(b) The classification error rate of $G_m(x)$ is calculated on the training data

$$e_m = P(G_m(x) \neq y_i) = \sum_{i=1}^{N} w_{mi} I(G_m(x) \neq y_i) \tag{4}$$

(c) Calculation the coefficient of $G_m(x)$, the logarithm is the natural logarithm

$$\alpha_m = \frac{1}{2} \log \frac{1 - e_m}{e_m} \tag{5}$$

(d) Update the weight value distribution of the training data

$$D_m = (w_{m+1,1}, \cdots, w_{m+1,i}, \cdots, w_{m+1,N}) \tag{6}$$

$$w_{m+1,i} = \frac{w_{mi}}{Z_m} \exp(-\alpha_m y_i G_m(x_i)) \tag{7}$$

where, Z_m is Normalization factor make the D_m become a probability Distributions.

$$Z_m = \sum_{i=1}^{N} w_{mi} \exp(-\alpha_m y_i G_m(x_i)) \tag{8}$$

(3) Build a linear combination of basic classifiers

$$f(x) = \sum_{m=1}^{M} \alpha_m G_m(x) \tag{9}$$

The final classification is

$$G(x) = sign(f(x)) = sign\left(\sum_{m=1}^{M} \alpha_m G_m(x)\right) \tag{10}$$

3 Driving Behavior Analysis System

The proposed driving behavior analysis system consists of driving operation acquisition module, data preprocessing module, the driving operation information fusion module, and AdaBoost classification and recognition modules.

Driving behavior analysis data will be divided into training set and test set. Preprocessing and feature extraction are simultaneously applied to both sets. Classification makes the test samples into the driving model based on AdaBoost algorithm to classify and determine the test sample category. The number of rightly or wrongly classified samples divided by the number of the test set samples is the classification correct rate or error rate.

The following example will use the driving data obtained from urban traffic road to illustrate the system processes. The processes for mountains and highway situations are similar. In the data acquisition module, at first, a good driving data and bad driving data should be collected as training set. Then collecting another data set includes good driving behavior data and bad driving behavior data as a test set using the same method. After data preprocess step, each time slice samples can be regarded as the rate of change the driving operation. This paper uses the training set to establish the driving classification model in the city as a judge model by the AdaBoost theory, and then the proposed system could use the test set to judgment the accuracy of the model. Finally, judge the merits of driving behavior. Fig. 2 shows the flowchart of the entire proposed system.

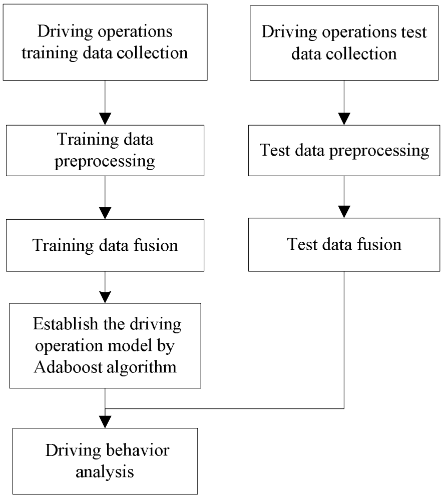

Driving operations training data collection
Driving operations test data collection
Training data preprocessing
Test data preprocessing
Training data fusion
Test data fusion
Establish the driving operation model by Adaboost algorithm
Driving behavior analysis

Fig. 2. The flowchart of the proposed driving behavior analysis system

4 Experimental Results and Analysis

4.1 Experimental Equipment

In this study, driving operation data are simulated through the YOYO driving training machine, produced by Guangzhou Great Gold East Network Technology Co. Driving operation data are read through the USB interface and combined with the simulation environment (city, mountains, and highway) to analysis the driving behavior. Fig. 3 shows the data collection equipment and driving simulated environment system.

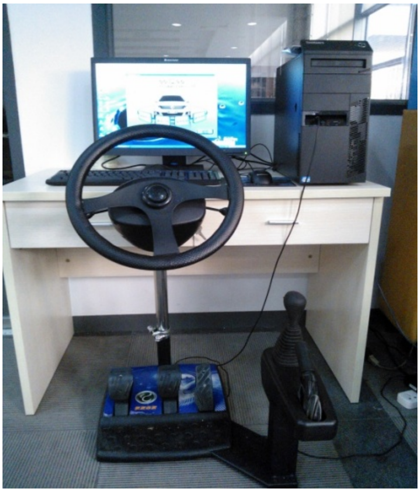


Fig. 3. Driving training machine

4.2 Data Collection

In the PC terminal, we used the data collection program (Bus Hound 6.01) real-time recording the driving operation data when driver driving simulation. Each record driving time is about 5-15 minutes and driving environments are selected from the city, highway, and mountainous road. Each of driving environment corresponds to simulate two driving data set, i.e., good driving behavior and bad driving behavior, respectively, for training set and testing set. The training set is used to establish driving model of city, highway, and mountainous road, respectively by AdaBoost. The test set is used to test validity about the above driving model established by AdaBoost. The method of collecting driving data is described as follows:

(1) Urban roads good driving record: maintain speed below 35 km/h, try to reduce the use of brake, deceleration in advance, try to keep a straight line driving, cornering speeds below 15 km/h, acceleration and braking are operated slowly. Bad driving record: Keep the car at speed driving state, quick step on and quick release the accelerator brake, remain uncertain direction and curve driving.

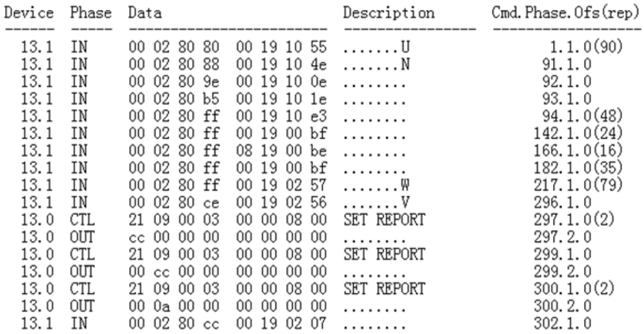
(2) Highway good driving record: maintain speed between 80 km/h and 100 km/h, keep the car is going along a straight line with almost no brake operation. No big turning in high-speed. Bad driving record: speed is too high and instability, curve driving, quick step on and quick release the accelerator brake, often braking.

(3) Mountain road good driving record: maintain speed below 40 km/h, turning speed below 20 km/h. Remained constant speed. Braking slowly and downshift in downhill. Bad driving record: speed is too high and instability, quick step on and quick release the accelerator brake, curve driving.

Finally, one can obtain the following 12 groups data with two sets.

4.3 Data Preprocessing

The data record format is shown in Fig. 4.



```
Device  Phase  Data                      Description       Cmd.Phase.Ofs(rep)
------  -----  ------------------------  ----------------  ------------------
  13.1  IN     00 02 80 80  00 19 10 55  .......U                  1.1.0(90)
  13.1  IN     00 02 80 88  00 19 10 4e  .......N                 91.1.0
  13.1  IN     00 02 80 9e  00 19 10 0e  ........                 92.1.0
  13.1  IN     00 02 80 b5  00 19 10 1e  ........                 93.1.0
  13.1  IN     00 02 80 ff  00 19 10 e3  ........                 94.1.0(48)
  13.1  IN     00 02 80 ff  00 19 00 bf  ........                142.1.0(24)
  13.1  IN     00 02 80 ff  08 19 00 be  ........                166.1.0(16)
  13.1  IN     00 02 80 ff  00 19 00 bf  ........                182.1.0(35)
  13.1  IN     00 02 80 ff  00 19 02 57  .......W                217.1.0(79)
  13.1  IN     00 02 80 ce  00 19 02 56  .......V                296.1.0
  13.0  CTL    21 09 00 03  00 00 08 00  SET REPORT              297.1.0(2)
  13.0  OUT    cc 00 00 00  00 00 00 00  ........                297.2.0
  13.0  CTL    21 09 00 03  00 00 08 00  SET REPORT              299.1.0
  13.0  OUT    00 cc 00 00  00 00 00 00  ........                299.2.0
  13.0  CTL    21 09 00 03  00 00 08 00  SET REPORT              300.1.0(2)
  13.0  OUT    00 0a 00 00  00 00 00 00  ........                300.2.0
  13.1  IN     00 02 80 cc  00 19 02 07  ........                302.1.0
```

Fig. 4. Schematic of collection data text

(1) Extract data packet from the "data" column of the Fig. 4. The packet is converted from hex to decimal. These data include steering, brakes, and accelerator information. The first column is the steering wheel angle information, expressed by the number from 0 to 255. The second column is the turn left and turn right information: Digital 1 represents the first circle turn to the left; 0 represents the second circle turn to the left; 2 represents the first circle turn to the right; 3 represents the second circle turn to the right. In order to facilitate the data analysis, the steering angle information will be converted into a continuous angle data from -510 to 510, negative numbers indicate turn left, positive number indicate turn right, Where each number represents 1.4118 degree angle. The third column is the brake and throttle Information, according the same method as changes of steering angle information, the throttle data is converted from -128 to 0, and the brake data is converted from 0 to 128.

(2) In the data record, the "Phase" column data indicates the data packet input or output state, "IN" represents the driving operation from driver training machine input information, "OUT, CTL" represents the pc control information output, we only need to extract the information that driver training machine input which the "IN" correspond to data packet from the "data" column.

(3) The time information processing, the "Cmd.Phase.Ofs (rep)" column represents the time series, where figures in brackets indicate the time of the operation remain unchanged. We will restore the driving operation information of each time slice and finally composite the driving operation record of the continuous time slice. The driving operation data of each time slice is a sample, including Attribute 1: steering wheel angle; Attribute 2: Brake throttle Information; and sample label, 1 represents a good driving, -1 represents a bad driving.

According to the above preprocessing methods, we get a series of graph about city, highway, or mountain road driving record data. Fig. 5 shows the city good driving record steering wheel data graph.

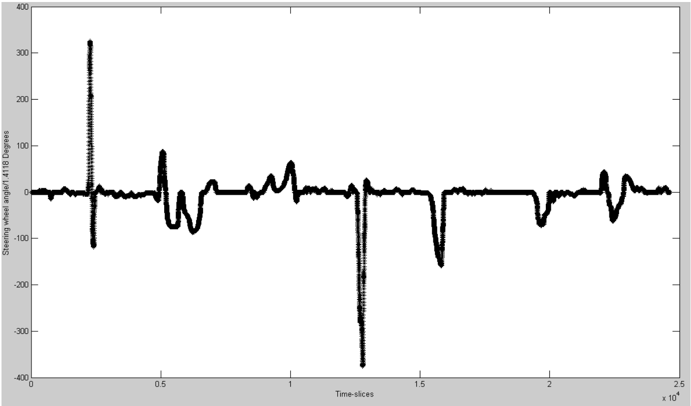


Fig. 5. The city good driving record data of steering wheel

4.4 Experiment

Feature Extraction

In this paper, the feature of every simple is choose the steering wheel angle change rate, the accelerator pedal change rate and the brake pedal change rate. The proposed system will combine these different traffic conditions to establish the driving model.

Experiment Result

In this paper, the experiments will use the Matlab tools box, called GML_AdaBoost_Matlab_Toolbox_0.3, to complete the driving modeling and test simulation. Table 1 shows the result of driving behavior analysis based on the AdaBoost algorithm.

Table 1. AdaBoost -based information fusion (steering wheel, brake throttle, road conditions) driving behavior analysis

	Good driving behavior correct rate	bad driving behavior correct rate	Comprehensive assessment
City road	95.95%	62.19%	79.07%
Highway road	93.69%	72.1%	82.90%
Mountain road	97.95%	45.69%	71.82%

4.5 Experimental Analysis

From the simulation results given in Table 1, one can clearly see where the testing process in the modeling. AdaBoost comprehensive test accuracy rate can reach 80%, where the good driving behavior recognition rate is relatively higher, while the identification of bad driving behavior is lower, that mainly caused by the following points: Since in this test we used the test and training data set which composed with multiple time-slice sample set include the steering wheel angle gradient and brake throttle gradient to establish the driving model in different road. The driving behavior analysis system will judge the each of the time slice sample in real time. While there are so many good driving behavior sample in the bad driving data set, such as Uniform motion in a straight, slow start and so on. That is the reason why the identification of bad driving behavior is lower.

5 Conclusion

With the real-time driving behavior analysis becomes more and more required, this paper utilizes a number of critical driving operation data (brakes, throttle, steering wheel angle and road conditions) to comprehensive analysis the driving behavior. This paper shown driving behavior analysis model based on AdaBoost can effectively achieve the correct judgment on driving behavior analysis and timely corrective driver's driving improperly.

Acknowledgement. This work was supported by University Innovation Project of Xiamen (3502Z20131158), National Natural Science Foundation of China (61273290), and the Ministry of Science and Technology, Taiwan (101-2221-E-366-002-MY3).

References

1. Leonard, E.: Traffic safety. Science Serving Society (2004)
2. Yan, X., Zhang, H., Wu, C., Mao, J., Hu, L.: Research progress and prospect of road traffic driving behavior. traffic information and safety (1), 45–51 (2013)
3. Miyaji, M., Danno, M., Oguri, K.: Analysis of driver behavior based on traffic incidents for driver monitor systems. In: IEEE Intelligent Vehicles Symposium, pp. 930–935 (2008)
4. Liang, J., Cheng, X., Chen, X.: The research of car rear-end warning model based on mas and behavior. In: IEEE Power Electronics and Intelligent Transportation System, pp. 305–309 (2008)
5. Wu, B., Ai, H., Huang, C., et al.: Fast rotation invariant multi-view face detection based on real AdaBoost. In: Proceedings of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004, pp. 79–84. IEEE (2004)
6. Li, H.: Statistical Learning Method. Tsinghua University Press (2012)
7. Parviz, M., Moin, M.S.: Boosting approach for score level fusion in multimodal biometrics based on AUC maximization. Journal of Information Hiding and Multimedia Signal Processing 2(1), 51–59

The Study and Application of IT Projects Management and Control Model for Enterprise Group

Ke Li and Yongteng Wang

CGN Power Co., Ltd., Centre of Information Technology, Shen Zhen, China
{like,wangyongteng}@cgnpc.com.cn

Abstract. In this paper, we are going to introduce PRINCE2 ,which is the advanced international projects management methodology, to build a model of projects management and control based on PRINCE2 & CMMI for enterprise group that has already been implementing CMMI in order both to manage and control the feasibility analysis, business justification, process monitor, resources management, appraisal of IT projects and ensure IT projects are necessary, feasible, economical, controllable, investment effective as well.

Keywords: PRINCE2, CMMI, Project Management & Control, Process Control.

1 Introduction

With the rapid development of information technology, many enterprise groups are integrating information technology and core business with management improvement in recent years. They are vigorously promoting the construction of information infrastructure construction. Therefore the actual input expenses of information technology and the number of IT projects are increasing rapidly. Accordingly, how to implement effective management and control on the numerous IT projects to make sure that each project can meet the initial expectation is a problem faced by many enterprise groups.

The introduction of IT project management international standard is the way that many enterprise groups are now adopting. For example, an enterprise establishes separately information technology management functions (The Information technology Functions) and professional IT service center (It Services Department) in the organizational structure of their headquarters.

It Service Department got the CMMI3 certificate in 2009. By putting CMMI standard into use, it can implement effective management on resources, quality, cost and schedule of IT projects.

But the project management, based on the CMMI's norm, is paying attention to manage a single project regularly and systematically and ensuring the whole project successfully through each link's management. Therefore this design is considering to the project manager. But what the information technology department confronted is a number of programs. Their core work is to control the programs and focus on the

H. Sun et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 329, DOI: 10.1007/978-3-319-12286-1_29

business case, scientific decision, process monitoring, resource management and project evaluation and so on. Only through the effective control and based on the project management to set up an effective group level of programs control system, can we ensure the IT item's scientific decision, to be correspond to the plan of group information technology, to increase the efficiency of project's implement, to decrease their project cost and ensure the investment benefits as well as to realize using information technology to support the development of enterprise.

This passage is through to study the managerial method of PRINCE2's item, to elaborate how to combine PRINCE2 with CMMI's norm, to establish an IT project management model which is suitable for a large enterprise groups and to implement the application.

2 The Circumstance of Application of CMMI

The capability maturity model provides a standard of norm and managerial requirement which not only can ensure the quality of products, but also can reduce the period of exploitation and improve their working efficiency. According to the CMMI's normative managerial requirement, the IT service department has accomplished to formulate and release 20 item's management procedure. At the same time, they ask for all the IT items must be carried out the projects in accordance with relevant standard of procedure in instituation. CMMI focuses on the importance of managing the process of the project's implement which is in order to play the important role of helping the project's manager to improve their managerial ability in a single project's lay.

3 PRINCE2 Project Management Methodology

PRINCE is short for Project IN Controlled Environment, which is paying attention to how to manage the projects in a logical and organized way in controlled environment and ensuring the project can be available to carry out within the controlled environment.

PRINCE2 is structured project management method which is adopting a method based on the process to control the projects. The process not only define management activities which is in the process of projects needed, but also describe the parts which the activity is included.

PRINCE2's management process consists of 8 special parts, covers almost all the project control and managerial activities from the beginning of the projects to the end of the projects. Including SU, DP, IP, CS, SB, MP, CP, PL the process of guidance and plan cover the whole activity. At the same time, the process of PRINCE2's management is based on 8 parts, including business case, organization, plan, control, risk management, quality, configuration management and changed control.

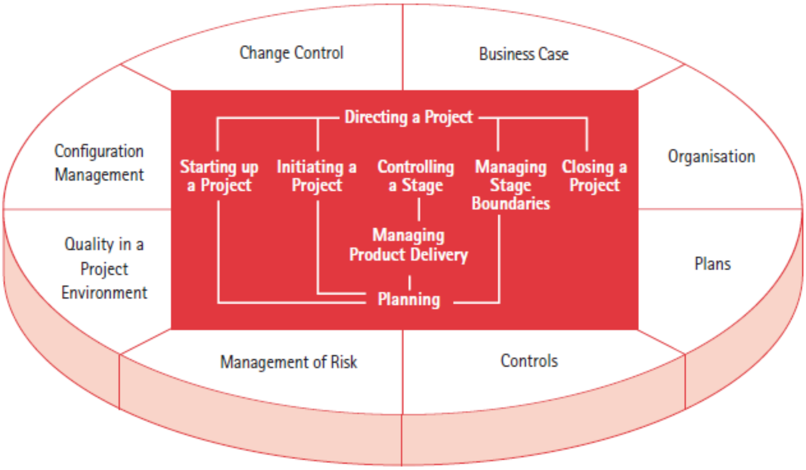


Fig. 1. PRINCE2 processes and components

Hence based on the frame of PRINCE2, and to combine the group of enterprise's actual condition, we have drowned up the process of PRINNCE2 and the corresponding content's relationships in a group of enterprise.

components processes	Business Case	Organisation	Planning	Controls	Management of Risk	Quality In A Project Environment	Configuration Management	Change Control
SU	√	√	√		√			
IP	√		√	√	√	√		
CS			√	√		√	√	√
MP			√	√		√		√
SB	√	√	√	√	√			
CP	√			√			√	
PL			√	√		√		
DP	√		√	√	√			

4 PRINCE2 & CMMI 's Projects of Group Control Model

4.1 The Comparison of PRINCE2 & CMMI

According to relevant standard of overlay analysis from 5 COBIT enterprise's management frame, as shown in the figure 2, on the life of the product span, CMMI only focuses on the project life cycle while the PRINCE2 cover more widely, including the former decision and later evaluation.

And the two parts is just what the information technology department concerned. Likewise CMMI is more careful and deeper than PRINCE to structure and implement the project in project management and this is also the IT department (particularly the project manager) is concerned. Hence how to use both projects better and accelerate their fusions are becoming a problem which need the IT project group of controls of enterprise to considerate and deal with.

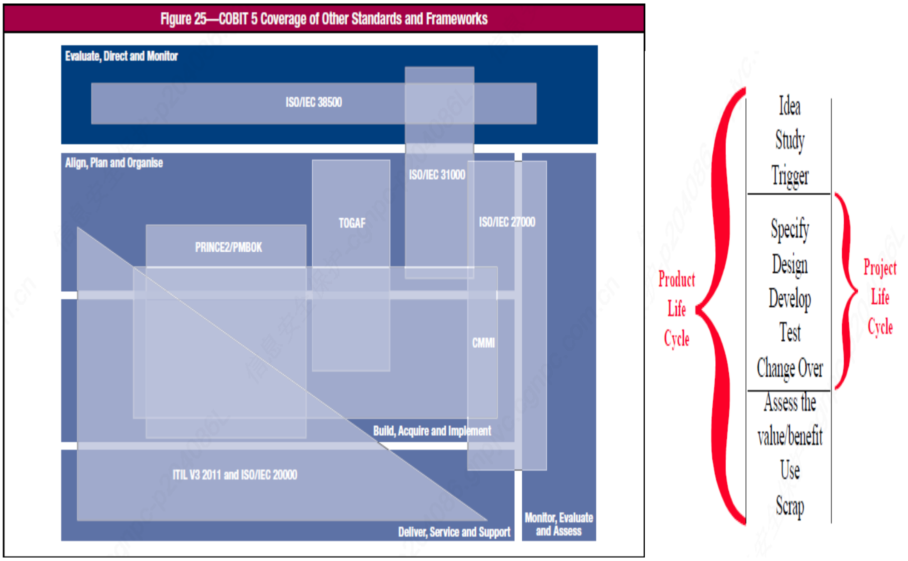


Fig. 2. Product Life Cycle

The information technology function sector and project guide give an analysis to project group of control level in one enterprise group(as shown in figure 3), the former at “Project Board” level which is in charge of group control job; IT service sector at “Project manager” level which is responsible for specific project management work.

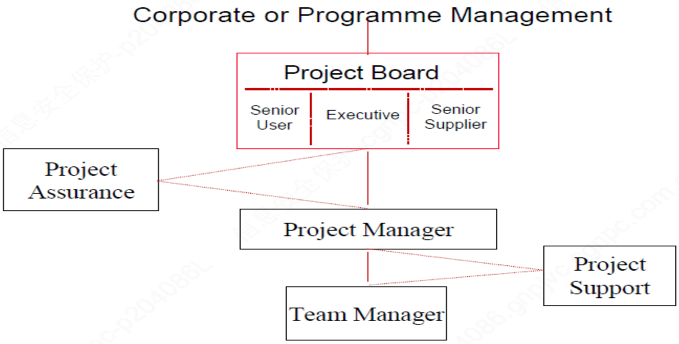


Fig. 3. Project management structure

In the process model of PRINCE2, “guiding project” level is realized by project guiding and five sub procedures, it will be throughout the whole project process from the beginning early stage of the project to continue to the end of the project post-evaluation(as shown in figure 4).

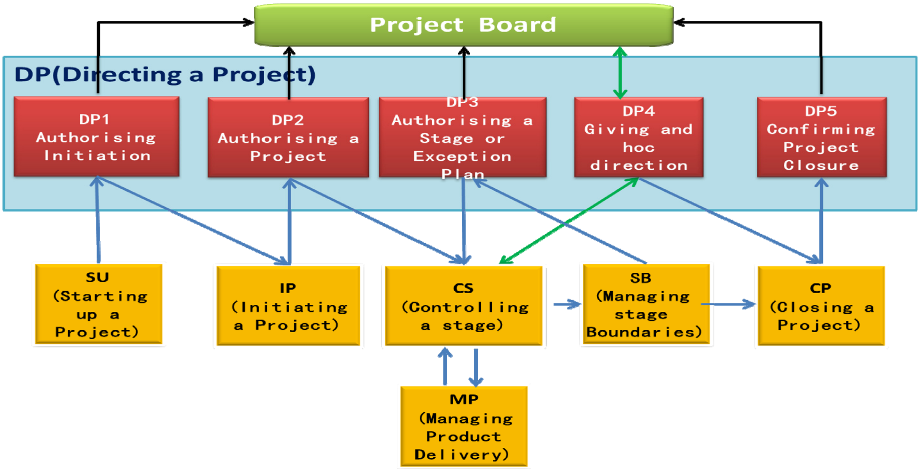

Project Board
DP(Directing a Project)
DP1 Authorising Initiation
DP2 Authorising a Project
DP3 Authorising a Stage or Exception Plan
DP4 Giving and hoc direction
DP5 Confirming Project Closure
SU (Starting up a Project)
IP (Initiating a Project)
CS (Controlling a stage)
SB (Managing stage Boundaries)
CP (Closing a Project)
MP (Managing Product Delivery)

Fig. 4. Directing a Project

DP1: Authorising Initiation: This procedure happens after the project preparation is done. The Information-based Functional Department, by consulting and analyzing the results that were attained in the preparation, will decide whether the project should be included in the plan or specific budget should be set aside. If it is passed, then come to the procedure of Start Project.

DP2 Authorising a Project: After starting the project, the Information-based Functional Department will organize experts to assess its business case, the matching rate between the project and the company’s strategies and plans, scheme and cost, risks and benefits, etc. They will decide whether to approve and initiate the project. If it is passed, then come to the procedure of Stage Control.

DP3 Authorising a Stage or Exception Plan: This procedure is introduced in when considerable changes (exceptive events) take place in any milestone of the construction or the scheme of the project. The Information-based Functional Department reviews the current situation and risk of the project, its matching rate to the scheme, resource demand and result target of the forthcoming stage, business case, prospective benefits, etc. , and decisions on whether to enter the next stage will be made on the basis of the conclusion. If the expectation exceeds the tolerant deviation which was given in the authorization, the Information-based Functional Department will consider and decide whether it should be reported to the higher authorities. With the approval of the higher authorities, the previous plan will be replaced by an exceptive plan. If the project doesn't deserve to be continued according to the decision, then the project manager should be informed to stop it and works off arrears orderly.

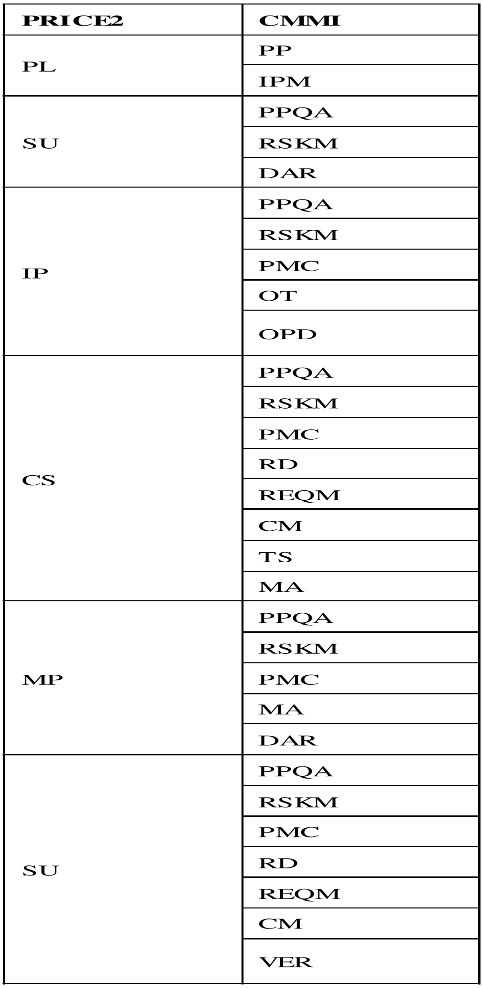
DP4 Giving and hoc direction: This procedure goes through the construction of the project. When the project is going on well as planned, the project manager can apply to the Information-based Functional Department for direction and coordination on the clarification of the program, the consideration of outer influence factors, resource allocation, inner conflicts and the alternation of project organization. When the project shows obvious deviation or a considerable change (an exceptive event) happens, the Information-based Functional Department should intervene actively, organizing reviews and giving a clear direction.

DP5 Confirming Project Closure: This procedure happens when all the project close-out has been properly done. The project manager submits the result of the project as well as the acceptance material for the Information-based Functional Department to organize appropriate party to check the project before acceptance and then decide the following schemes and finish the sum-up report.

By analyzing the correspondence between the procedures and contents, the class that the Information-based Functional Department belongs to and the related sub-procedures of the Project Direction, we can see what the Information-based Functional Department does in the management and control of the project groups coincides with the process of the Project Preparation in PRINCE2, covering and proceeding through the whole project construction circle, from Project Preparation, Starting Project, Stage Control to Boundary and Close-out Supervise.

4.2 IT Service Department and Project Management

In the project construction circle, PRINCE2 also has clear and specific requirements and directions. Considering IT Service Department has passed through the recognition of CMMI3, lain down project management program which satisfies CMMI, and carried out direction. These several years' operation has proved it is practical and effective. The requirements and directions of PRINCE2 on the project management can be an equivalent to the procedures of present CMMI management.



PRICE2	CMMI
PL	PP
	IPM
SU	PPQA
	RSKM
	DAR
IP	PPQA
	RSKM
	PMC
	OT
	OPD
CS	PPQA
	RSKM
	PMC
	RD
	REQM
	CM
	TS
	MA
MP	PPQA
	RSKM
	PMC
	MA
	DAR
SU	PPQA
	RSKM
	PMC
	RD
	REQM
	CM
	VER

4.3 IT Project Control Model

Through the study of the analysis of PRINCE2 process, we combine with the characteristics of the information technology department in an enterprise group which is responsible for the guidance and control project, and the characteristics of the IT services department which is responsible for the project construction and management, considering the situation of the established CMMI specification, the IT project group control model is studied and established based on PRINCE2 and CMMI.

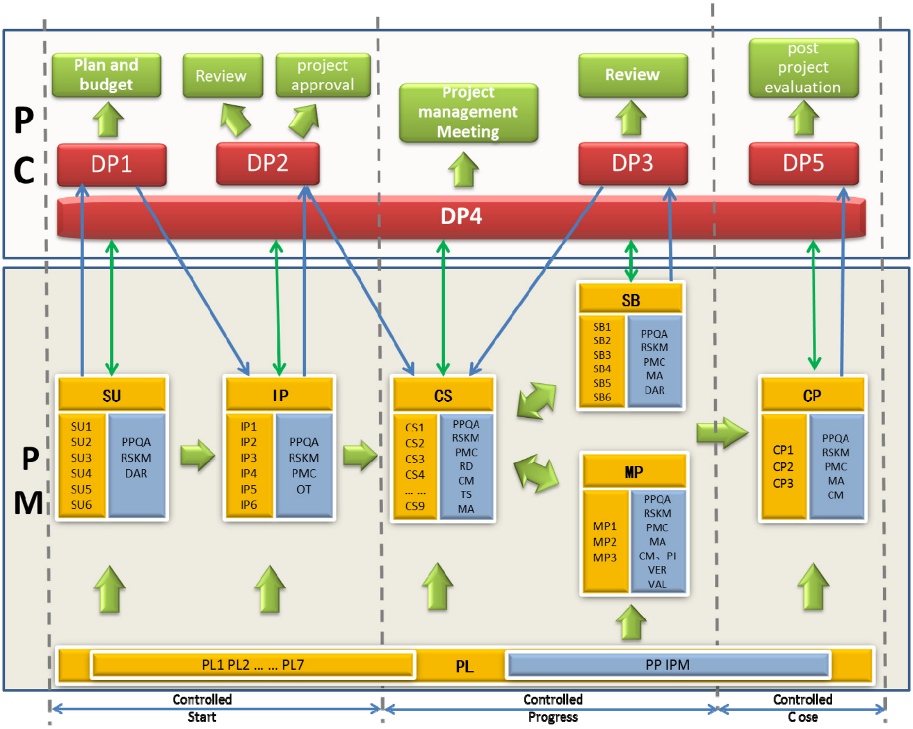


Fig. 5. IT project group control model is studied and established based on PRINCE2 and CMMI

4.4 Controlled Start

The users give projects. The project manager will research the feasibility and then make a feasibility report. After that the report will be sent to the information departments for review.

Combining the information technology planning and annual work plan of the enterprise, the information departments will examine and verify the feasibility reports, confirm the annual IT project plan and budget.

Through the budget audit, the construction scheme of the project and technical specification will be made, and then the information departments will organize to review

The information departments organize experts to the group level special review of the project. After the special review, the project comes up to the project approval and the examination.

The project manager can make an application to the information departments to guide and coordinate, to inspect the project and give specific suggestions for improvement when he works on the clarification of the project, the allocation of resources or when he meets internal conflict and project organization changes.

4.5 Controlled Process

According to the requirements of the enterprise IT project management, the project manager manages the construction of the project according to the plan and stages.

According to the plan, the project manager makes the stages of review of the process and results of each stage of project activities.

In the course of construction of the project, the project manager needs to report to the information departments in every important milepost stage or plan changes (exceptions).The information departments evaluate the status of the project and make decision including approving exception plans to the next stage, notifying the project manager to terminate the project or report to the project leadership team.

The project manager can make an application to the information departments to guide and coordinate when he works on the allocation of resources or when he meets internal conflict and project organization changes. The information departments organize to inspect the project and give specific suggestions for improvement when there is an exception or a serious deviation.

4.6 Controlled Ending

The project manager completes project acceptance stage activities in accordance with the requirements of project management, submits the results and materials of acceptance to the information department

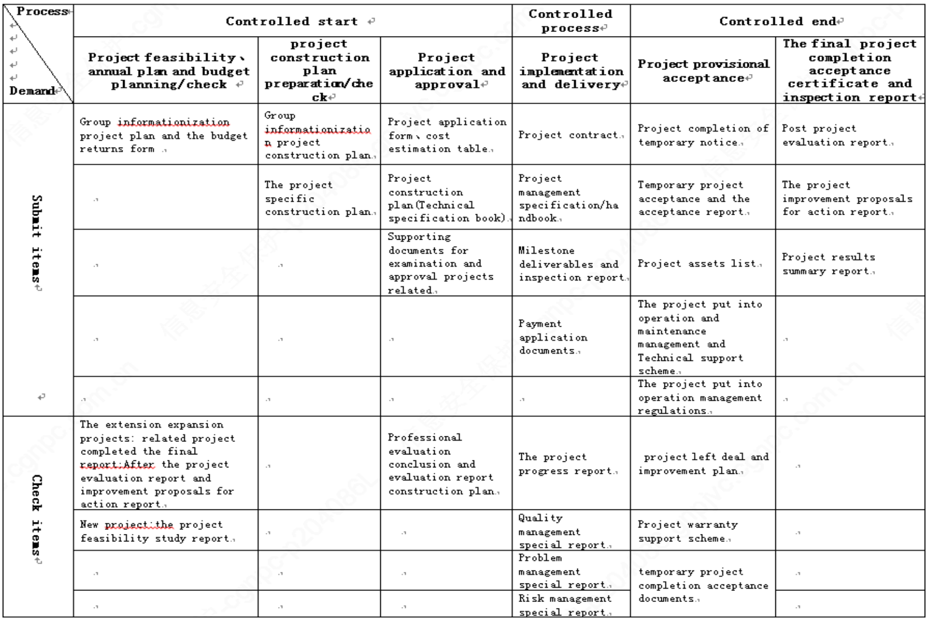
Information departments organize the relative parties to check up the project completion and acceptance, to determine follow-up action list, to complete the summary report, to arrange the evaluation for the investment value and benefit of the project.

The project manager can make an application to the information departments to guide and coordinate if problems happen in the course of the project acceptance. Or the information departments organize to inspect the project and give specific suggestions for improvement when there is an exception or a serious deviation.

4.7 Preliminary Application Model

To make project stakeholders clearly understand the project control requirements of the information technology department in the model, IT project construction process and management requirements of an enterprise group are established.

For timely tracking and understanding all the group IT projects of the enterprise, based on the control model, the information function department establishes an IT project planning and budget tracking system in an enterprise group, project evaluation system and project processes (see figure 6). Each project manager feedbacks the progress of IT projects of the above system, and realizes IT online examination and approval and control of the project in the enterprise.

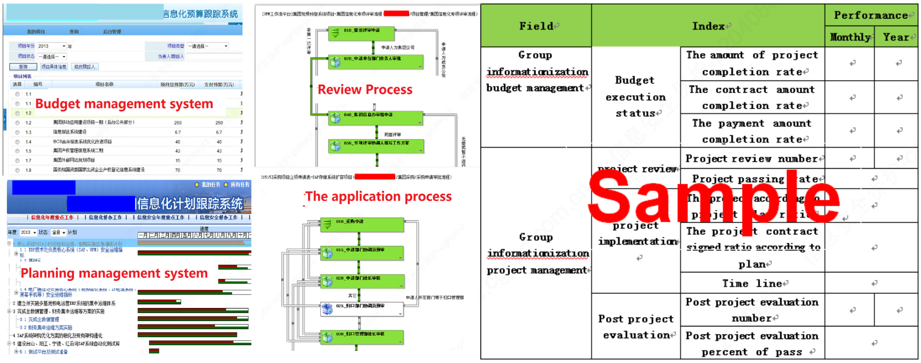


Process / Demand	Controlled start			Controlled process	Controlled end	
	Project feasibility、annual plan and budget planning/check	project construction plan preparation/check	Project application and approval	Project implementation and delivery	Project provisional acceptance	The final project completion acceptance certificate and inspection report
Submit items	Group informationization project plan and the budget returns form	Group informationization project construction plan.	Project application form、cost estimation table.	Project contract.	Project completion of temporary notice.	Post project evaluation report.
		The project specific construction plan.	Project construction plan(Technical specification book).	Project management specification/handbook.	Temporary project acceptance and the acceptance report.	The project improvement proposals for action report.
			Supporting documents for examination and approval projects related.	Milestone deliverables and inspection report.	Project assets list.	Project results summary report.
				Payment application documents.	The project put into operation and maintenance management and Technical support scheme.	
					The project put into operation management regulations.	
Check items	The extension expansion projects: related project completed the final report;After the project evaluation report and improvement proposals for action report.		Professional evaluation conclusion and evaluation report construction plan.	The project progress report.	project left deal and improvement plan.	
	New project:the project feasibility study report.			Quality management special report.	Project warranty support scheme.	
				Problem management special report.	temporary project completion acceptance documents.	
				Risk management special report.		

Fig. 6. IT project construction process and management requirements of an enterprise group

Meanwhile, when there is a special exception in the IT projects, for example, a serious progress deviation occurs in the intelligent project of a building (more than two months behind schedule), the function department of the information technology, in accordance with the requirements of DP4 process, organizes project leading group meetings and carries out the specific improvement action, making sure the current project is gradually catch up with the original schedule.

To better apply the project control model, and help implement enterprise group real-time monitor and manage the project group, the function department of the information technology plans the next step in implementing the requirements of the control mode, and meanwhile develops project control performance indicators, and establishes a system for real-time monitoring indicators, IT projects can ensure the enterprise group is in a controlled environment construction and completed smoothly.



Field	Index		Performance	
			Monthly	Year
Group informationization budget management	Budget execution status	The amount of project completion rate		
		The contract amount completion rate		
		The payment amount completion rate		
Group informationization project management	project review	Project review number		
		Project passing rate		
	project implementation	The project according to project [illegible]		
		The project contract signed ratio according to plan		
		Time line		
	Post project evaluation	Post project evaluation number		
		Post project evaluation percent of pass		

5 Conclusion

By establishing the model of IT project group control based on PRINCE2 and CMMI, the information functions don't have to focus on the concrete construction process of each project. What we have to do is to concentrate on the control of project group and the management by exception on specific items. Theoretically, in this way, we can complete a series of controllable, visible and well-managed activities and hence achieve the desired aims under the guidance of information intelligence department.

After the practical application of this model as well as part of the control system, we have achieved the preliminary track and control of IT project throughout the annual process.

The information intelligence functions has strengthened the audit efforts of 2013 IT Project Business Case, the yearly budget of IT project decreased about 20% year-on-year.

In the link of project authorization (DP2) ,in the first half of 2013, IT projects experienced 33 assessments, and the one-time pass rate is 85%, and the coverage of the project was as wide as 100%.

During the process or in the link of exception plan (DP3) and special guide (DP4), we enhanced the check and guidance of the annual key projects, such as intelligence construction of buildings, management and decision support system as well as the consolidated financial statements. We also organized more than ten meetings on different management levels and solved a number of major issues over the construction process of IT project. .

In the confirmation part of the project closeout, the evaluation of ERP construction project (ERP system Status Diagnostic Evaluation) is completed, which is the key project of "Eleventh Five-Year Plan" of group information technology.

It is noted that the group control model needs constant improvement and optimization according to the practical application effect, and thereby realize ultimately the Process, indexation, systematization and real-time transformation of IT group project.

References

[1] OGC. Managing Successful Projects with PRINCE2. The Stationery Office (2009)

[2] Bentley, C.: Prince2: A Practical Handbook. Butterworth-Heinemann Ltd. (2001)

[3] Ahern, D.M., Clouse, A., Turner, R.: CMMI Essence, 3rd edn. Tsinghua University Press, Beijing (2009)

[4] ISACA. A Business Framework for the Governance and Management of Enterprise IT (2012)

[5] Hinde, D.: PRINCE2 Study Guide. Wiley (2012)

[6] Graham, N.: PRINCE2 For Dummies (2010)

[7] Project Management institute, A Guide to the Project Management Body of Knowledge (PMBOK Guide) (2013)

[8] Hughes, B., Cotterell, M.: Software Project Management, 5th edn. (2010)

Part V
Intelligent Multimedia Tools and Applications

A Data Hiding Method for Image Retargeting

Wen-Pinn Fang, Wen-Chi Peng, Yu-Jui Hu, Chun Li, and Shang-Kuan Chen

Department of Computer Science and Information Engineering, Yuanpei University,
300 HsinChu, Taiwan
wpfang@mail.ypu.edu.tw

Abstract. Now-a-day, it is necessary to modify image size for fitting different display devices. This process is called image retargeting. If users hide a secret image into a host image. The hiding data may loss after the retargeting process. This paper proposed a data hiding method which the secret image will not loss after the image has been retargeted. Based on the fault-tolerance property of secret image sharing scheme, many shares are generated from a secret image. The shares are embedded in many locations of the host image. This means the hiding data may not delete after the image has been retargeted. The result is suitable for the video copyright protection between different devices. An experiment is also presented.

Keywords: Data hiding, image retargeting, secret sharing.

1 Introduction

Recently, there are many products are invented to show videos and images. The products include television, smart phone, tablet, and computer. The resolutions of these devices are different. Because cost consideration, there is only one version in the produce step every movie. It is necessary to resize the video for many display devices. Furthermore, it is very important to protect the producer's copy right. There are many approaches to protect the copy right. For example, digital signature, watermarking, data hiding*...etc.al*. However, the hiding data may disappear after image processing. This paper studies a method to overcome this problem especially focus on cropping type image retargeting.

The rest of this paper is organized as follows: the background knowledge is show in Section 2; the method is proposed in Section 3; Experimental results are shown in Section 4. Finally, the discussion is represented in Section 5.

2 Background Knowledge

Before describe the method, it is necessary to explain some basic knowledge. The basic knowledge include secret image sharing, image retargeting and data hiding. The detail is show as below:

H. Sun et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 329, DOI: 10.1007/978-3-319-12286-1_30

2.1 Secret Image Sharing

Secret sharing was first introduced by Shamir [1]. It is a reliable method for the protection of cryptographic key with many good properties. It is a perfect threshold scheme, with the size of each share not exceeding the size of the secret and the security does not rely on unproven mathematical assumptions. It is presented below as mention in Ref [2]:

Initialization Phase

1. D choose w distinct, non-zero elements of **Z**p, denoted x_i, $1 \le i \le w$(this is where we require p≥w+1). For$1 \le i \le w$, D gives the value xi to Pi. The values xi are public.

Share Distribution

2. Suppose D wants to share a key K∈ **Z**p.D secretly chooses (indepently at random) t-1 elements of **Z**p, $a_1, \ldots a_{t-1}$.
3. For$1 \le i \le w$,D computes $y_i = a(x_i)$, where

$$a(x) = K + \sum_{j=1}^{t-1} a_j x^j \mod p.$$

4. For $1 \le i \le w$,D gives the share yi to P_i

Fig. 1. The Shamir (t-w)-threshold scheme in $\mathbf{Z}_\mathbf{p}$

In 2002, Thien and Lin extend the scheme to image [3], named "Secret Image Sharing". They change the values of a_j into a corresponding pixel value of a secret image. In an (r, n) image sharing system [1][4][5], n shares {L_1, L_2,..., L_n} are created for a given image, e.g., Lena. The image can be revealed when all n shares are received, while less than n shares reveal nothing about the image. Sharing is a safety process that is valuable in a company where no employee/investor alone should be trusted. Significantly, the original image can be discarded after the sharing; moreover, each of the n shares is $1/n$ of the size of the given image. Therefore, the sharing process does not waste storage space.

2.2 Image Retargeting

In the decade, many researchers discuss the method of video resizing. The methods include image cropping style [6-9], seam carving [10-14], wrapping [15-17] and hybrid approaches [15-17]. The cropping method reserved the most important region directly. Avidn and Shamir [10] proposed a method to find out the seams first, remove the unimportant region, reserved significant regions. Many researchers extended the study. For example, Mansfield, *etc. al*[8], Rubinstein and Shamir[12], Grundmann, *etc. al*[13] proposed some suggestions to improve the quality and eliminated errors. On the other hand Wan[16] and Li[17] discuss the deformation effect after an image has been retargeted. Rubinstein [18] combined cropping and zooming method to achieve resizing. Sun, *etc. al*[19] proposed cyclic seam caving algorithm for the same purpose. Dong, *etc. al* [20] discussed the relationship between reservation region and whole result. Pritch [21] study the deformation and cropping relationship by shift-map method, the paper get better image quality. Ding [22] designed specific filter and get better image quality after the image has been retargeted.

2.3 Data Hiding

In general, there are three types of data hiding: vector quantization、error expansion and reversible displacement. The advantages of vector quantization type hiding method are compress and information hiding. However it's capacity is less than most other methods and with low image quality. The Error expansion type data hiding method has high storage capacity but the image quality is much lower. The complexity of reversible displacement type data hiding method is small and the quality of image is high. However, the hiding rate is smaller than other methods.

3 Proposed Method

The goal of this paper is design a data hiding method which the secret image will not disappear after the image has been retargeted. The symbol is show as Table 1.There are two phases in the method: (a) hiding phase (b) recovery phase.

In hiding phase, as shown in Fig.2, there are two steps:

Step 1: encode secret image into shares.

Using the (n,r) secret sharing scheme as shown in section 2.1.

For example, if the size of Host image is a gray scale image which the size is 512 by 512 pixels, the size of secret is a binnary image which the size is 64 by 64. The sharing scheme is (2621441,4096) sharing.

Step 2:hide the shares and pivots (the sharing parametric) into the host image. Here, the pivot means the location in the host image which store the relative coefficient of secret sharing scheme.

The hiding method is shown as below:

If the least two significant bits of pixel value of host image is l_1 and l_2. The hiding value of secret image is 1 if l_1 is equal to l_2, otherwise the value is 0 and it is the nearest number compare with original value. Before hide the secret image, the method hide the coefficient of secret sharing scheme in the left top most skin color pixel. Because the method never remove skin color pixel. Users can decode the shares. After hide the decode coefficients, scan the whole host image to hide share data pixel by pixel.

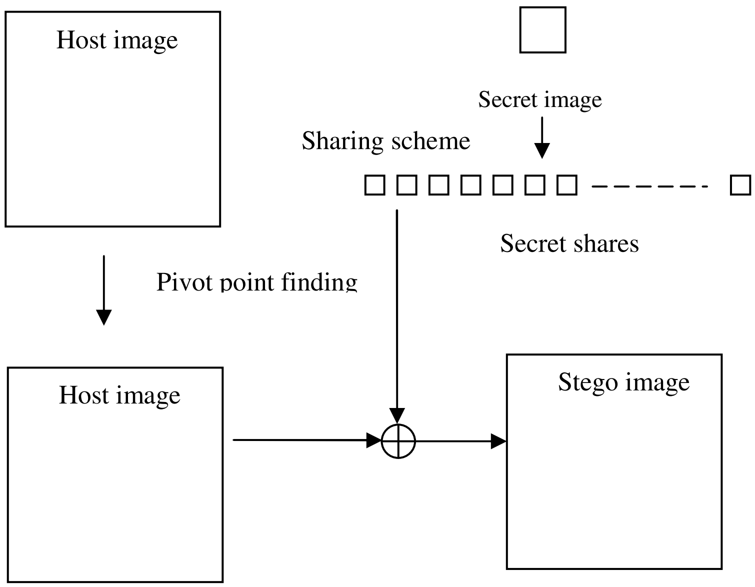


Fig. 2. The proposed hiding method

In recovery phase, there are three steps:

Step 1: retrieve the sharing coefficient from pivot.

Step 2: retrieve shares from stego image.

The value of shares is equal to $s_i(x,y)$

$$s_i(x,y)=P_{(x,y)}(l_1) \text{ xor } P_{(x,y)}(l_2) \tag{1}$$

Step 3: recovery secret image.

Recover secret image by inverse matrix.

Table 1. Font sizes of headings. Table captions should always be positioned *above* the tables.

Symbolic	description
H	Host image
S	Secret image
s_i	The ith share
$p(x,y)$	The pixel value in location (x,y)

For example, assume that the size of host image is 512-by-512 pixels. The depth of host image is 8 bits. The size of secret image is 64-by-64 pixels. It is a binary image. In the first step, scan the secret image pixel by pixel. Splitting the pixels into 8 bits non-overlapped sectors. Combine every sector into an integer which the value range is from 0 to 255. Let the numbers are p_0, p_1, p_2,…., p_{511} In the second step, design a polynomial

$$f(x)=p_0+p_1x+p_2x^2+....+p_{511}x^{511} \tag{2}$$

In the third step, plug in x from 1 to 4096, there will be 4096 numbers, notice, all operator are in GF(2^8).

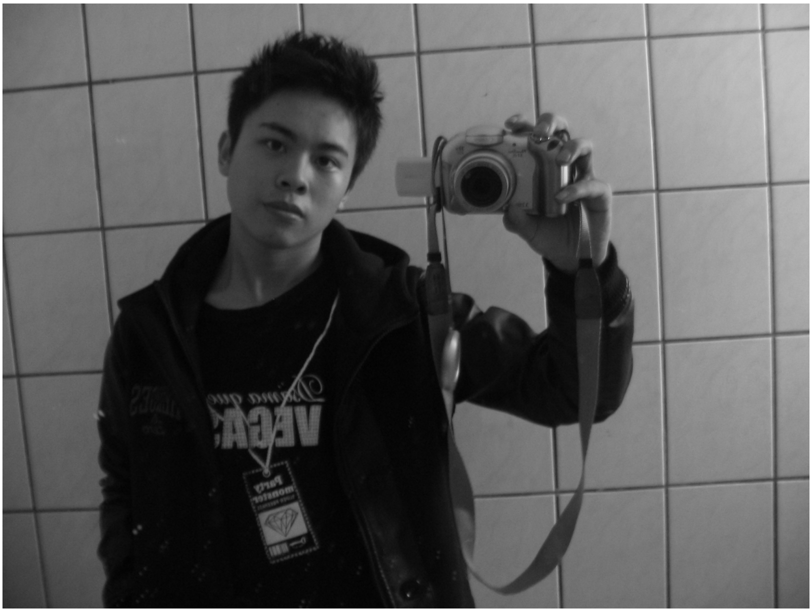
In the fourth step, split the 4096 number into 32768 bits by scanning the bit plane.

In the fifth step, replace the two least significant bits of host by equation (1).

4 Experimental Results

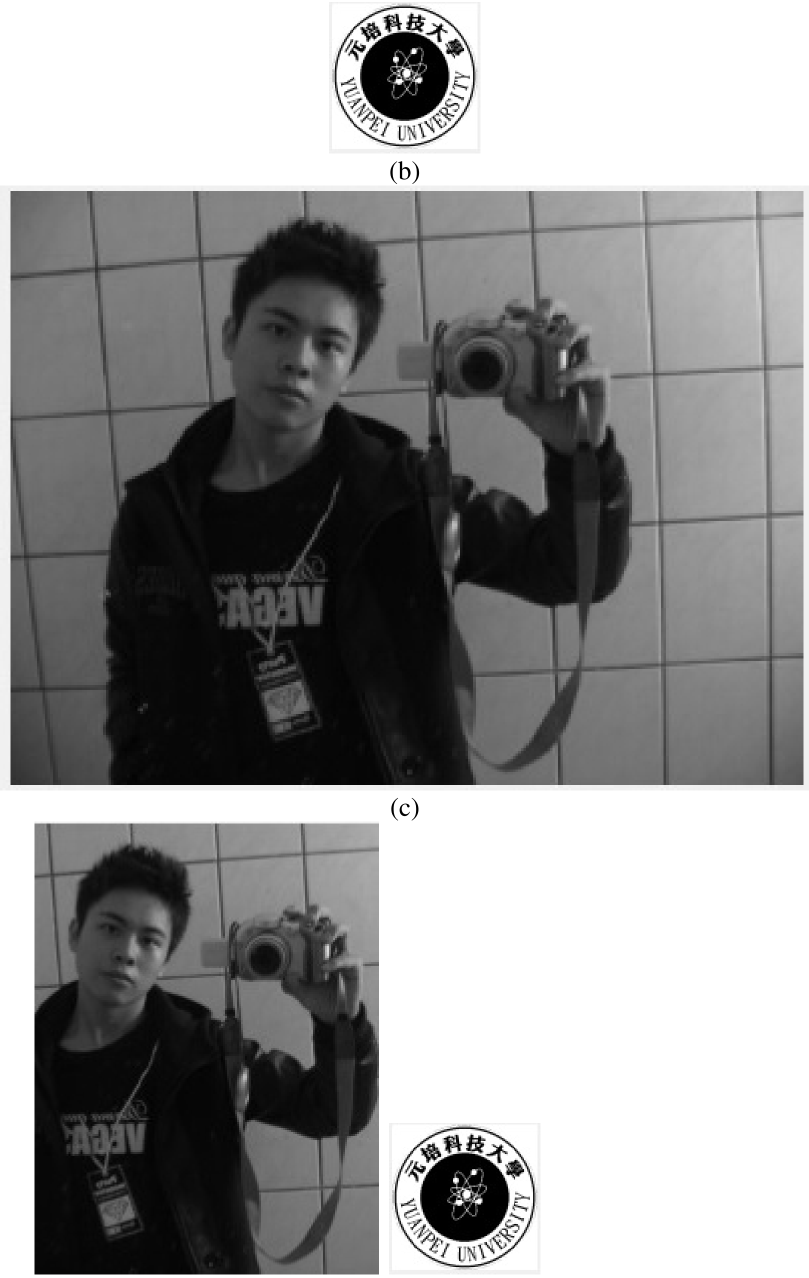
In this section, an experiment has been tested. There are three steps to test the proposed method. First at all, a secret image has been hided into a host image by the proposed method. Secondly, the stego image has been retargeted. In the third step, retrieve the secret image. Finally, compare the original secret image and the recovery secret image. Check the different between stego image and host image.

The test retargeting method is a cropping method. There are three steps: (a) find the most important part of image. In this testing case, the important part is the location of human. (b) cut the original image to fit the scale of display device. In this testing case is remove the background without people. The experimental result is shown in Fig.3. In this case, the secret image is the same as recovery image. After a series of experiments, The PSNR between host image and stego image is more than 35 dB.



(a)

Fig. 3. The experimental result, (a) is the original image (b) is the secret image (c) is the stego image (d) is the image after retargeting (e) is the recovery image



(b)

(c)

(d)

(e)

Fig. 3. (*continued*)

5 Conclusion and Remark

Image retargeting is a very important image processing method today. There are many display devices which the sizes are different. To make user enjoy the multimedia, image retargeting is necessary. However, it is difficult to preserve the hiding data after it has been retargeted. This paper proposed a data hiding method to overcome the problem by cropping type retargeting method. In the future, maybe use another hiding method

References

1. Shamir, A.: How to share a secret. CACM 22(11), 612–613 (1979)
2. Stinson, D.R.: Cryptography Theory and Practice, p. 327. CRC, USA (1995)
3. Thien, C.C., Lin, J.C.: Secret Image Sharing. Computers & Graphics 26(5), 765–770 (2002)
4. Thien, C.C., Lin, J.C.: An Image-Sharing Method with User-Friendly Shadow Images. IEEE Transactions on Circuits and Systems for Video Technology 13(12), 1161–1169 (2003)
5. [3] Thein, C.C., Fang, W.P., Lin, J.C.: Sharing Secret Images by Using Base-transform and Small-size Host images. International Journal of Computer Science and Network Security 6(7), 219–225 (2006)
6. Chen, L., Xie, X., Fan, X., Ma, W., Zhang, H., Zhou, H.: A visual attention model for adapting images on small displays. ACM Multimedia System Journal 9(4), 353–364 (2004)
7. Liu, H., Xie, X., Ma, W.-Y., Zhang, H.: Automatic browsing of large pictures on mobile devices. In: Proceedings of ACM International Conference on Multimedia, pp. 148–155 (2003)
8. Suh, B., Ling, H., Bederson, B., Jacobs, D.: Automatic browsing of large pictures on mobile devices. In: Proceedings of UIST, pp. 95–104 (2003)
9. Santella, A., Agrawala, M., Decarlo, D., Salesin, D., Cohen, M.: Gaze-based interaction for semiautomatic photo cropping. In: Proceedings of CHI, pp. 771–780 (2006)
10. Avidan, S., Shamir, A.: Seam carving for content-aware image resizing. ACM Transaction on Graphics 26(3), Article 10 (2007)
11. Mansfield, A., Gehler, P., Van Gool, L., Rother, C.: Scene carving: Scene consistent image retargeting. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part I. LNCS, vol. 6311, pp. 143–156. Springer, Heidelberg (2010)
12. Rubinstein, M., Shamir, A., Avidan, S.: Improved seam carving for video retargeting. ACM Transaction on Graphics 27(3), Article 16 (2008)
13. Grundmann, M., Kwatra, V., Han, M., Essa, I.: Discontinuous seam carving for video retargeting. In: Proceedings of International Conference on Computer Vision (2010)
14. Liang, Y., Su, Z., Luo, X.: Patchwise scaling method for content-aware image resizing. Signal Processing 92(5), 1243–1257 (2012)
15. Gal, R., Sorkine, O., Cohen-or, D.: Feature-aware texturing. Computer 11(5), 1–7 (2006)
16. Wang, Y., Tai, C., Sorkine, O., Lee, T.: Optimized scale-and-stretch for image resizing. ACM Transaction on Graphics 27(5), Article 118 (2008)

17. Li, B., Chen, Y., Wang, J., Duan, L., Gao, W.: Fast retargeting with adaptive grid optimization. In: Proceedings of International Conference on Multimedia & Expo, pp. 1–4 (2011)
18. Rubinstein, M., Shamir, A., Avidan, S.: Multi-operator media retargeting. ACM Transaction on Graphics 28(3), Article 23 (2009)
19. Sun, J., Ling, H.: Scale and object aware image retargeting for thumbnail browsing. In: Proceedings of International Conference on Computer Vision (2011)
20. Dong, W., Zhou, N., Paul, J., Zhang, X.: Optimized image resizing using seam carving and scaling. In: Proceedings of ACM SIGGRAPH Asia 2009, vol. 28 (2009)
21. Pritch, Y., Kav-Venaki, E., Peleg, S.: Shift-map image editing. In: Proceedings of International Conference on Computer Vision, pp. 151–158 (2009)
22. Ding, Y., Xiao, J., Yu, J.: Importance filtering for image retargeting. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (2008)

Rendering 3D Solid Model

Cheng-Wei Huang[1], Ran-Zan Wang[1], Shang-Kuan Chen[2,*], and Wen-Pin Fang[2]

[1] Department of Computer Science and Engineering, Yuan-Ze University
[2] Department of Computer Science and Information Engineering, Yuan-Pei University
skchen@mail.ypu.edu.tw

Abstract. A method of filling 3D model using generated solid textures is proposed. Using 3D texture synthesis, a 2D exemplar texture is synthesized to be a solid texture. The synthesized solid texture can be the stuffed material of 3D Solid model or the solid material of surficial mesh. The proposed method designs two algorithms to cutting the model. After cutting, we will repair the cross section. It makes the surface model look like the solid model. And it is real time display when cutting. Because, we can preview the 3D model and it internal cross section before printing, the application can also be combined with 3D printing.

Keywords: texture synthesis, solid texture, models cutting, 3D printer.

1 Introduction

Recently, texture synthesis is a popular technique for 2D or 3D texture pasting. The most classical procedural texturing method is the noise function proposed by Perlin[1]. This kind of method does not require any texture; however, it generate 2D or 3D image by only computation. The method is most used for simulating nature textures, for example, the ripple of water, a blaze of fire, or rocks. The drawback of this kind of method is too complicated for common users.

The examplar-base texture synthesis method reduces the procedures for users. It adopts a texture with highly replication for synthesizing an image. The exemplar-base texture synthesis method can be categorized into two cases. The one case is based on vertex modification [2-4]. For each vertex, the basic idea is to find the most similar blocks from the neighboring blocks of x-y, y-z, and z-x plane, respectively, and then replace the middle vertex with the color of the vertex. Because the method is executed vertex-by-vertex, it is time-consuming. The time complexity is $O(N_s \times N_e \times n)$, where N_s and N_e be the numbers of input vertices and output vertices, respectively, and n be the numbers of vertices in three neighboring planes. The other case is based on block modification. In 2013, Du et.al [5] proposed a method that stacks distinct particles to produce 3D solid texture. First, a mask image is generated by using segmenting, region growing to apply on an input 2D texture. Then, use the input 2D texture and the generated mask image to find the x-y, y-z, z-x cross sections of the particle. Later, the

[*] Corresponding author.

H. Sun et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 329, DOI: 10.1007/978-3-319-12286-1_31

shape and the color of the particle can be obtained by interpolation. Based on the cross section of the input texture and the related location between cross sections, a complete solid texture is generated. Finally, adjust the location and the size of each particle to be optimized.

Takeayama et. al [6] proposed a method called "Lapped Solid Texture" in 2008. The method adopts solid texture as the input. Because the input solid textures are with different characteristics in inner and outer parts, it cannot use 3D texture mapping for filling the mesh. The method needs to customize some solid textures for the input. Another topic of [6] is model cutting. User can draw any line on the model and then cut the model to see inner structure.

The remainders of this paper are organized as follows: The proposed method is stated in Section 2. The experiment results are shown in Section 3, and a conclusion is summarized finally in Section 4.

2 The Proposed Method

The proposed 3D model filling method using generated solid texture is categorized into two stages: (1) solid texture synthesizing, and (2) model cutting. In the stage of solid texture synthesizing, a 2D texture is synthesized to be a 3D texture with similar structure. In the stage of model cutting, users can immediately see the filled texture of inner structure of the model.

2.1 Solid Texture Synthesizing

In this subsection, the proposed solid texture synthesizing method can be divided into two main phases: searching and optimizing. Finally, the histogram mapping proposed by Kopf et.al [2] is used for resembling the color distribution with the original image.

In the searching phase, for each vertex of output solid texture, the most similar vertex in the input texture will be found by the following rule. For each vertex of output solid texture, there are three target planes perpendicular to each other. For the neighboring block of the vertex on the target plane, the rule is to search the vertex with most similar neighboring block in the input texture. First, a global energy function is defined as follows.

$$E(s,e) = \sum_{v\in V} \sum_{i\in\{x,y,z\}} \sum_{u\subset N_i(v)} \left\| s_{v,i,u} - e_{v,i,u} \right\|^r , \quad (1)$$

where s and e represent the output solid texture and input texture image, respectively, V represents the set of vertices, x, y, z represent the neighboring blocks of X, Y, Z planes, respectively, $N_i(v)$ represents the vertices of neighboring block, and $s_{v,i,u}$ represents the neighboring vertex of plane i of the s_v in the solid texture.

In the optimizing phase, the more suitable color will be computed by using the mapping between the vertex of neighboring block and the color information of input texture. Formula (1) is rewritten as follows.

$$E(s,e) = \sum_{v\in V} \sum_{i\in\{x,y,z\}} \sum_{u\subset N_i(v)} \omega_{v,i,u} \left(s_{v,i,u} - e_{v,i,u}\right)^2 \quad , \tag{2}$$

where $\omega_{v,i,u} = \| s_{v,i,u} - e_{v,i,u} \|$ is the weight of each vertex corresponds to medium vertex s_v. To compute the suitable s_v, the following local energy function can be adopted.

$$s_v = \frac{\sum_{i\in\{x,y,z\}} \sum_{u\in N_i(v)} \omega_{u,i,v} e_{u,i,v}}{\sum_{i\in\{x,y,z\}} \sum_{u\in N_i(v)} \omega_{u,i,v}} \tag{3}$$

In the histogram mapping phase, an algorithm is designed for solving the color unbalance problem. Firstly, the histogram is divided into 16 bins. For each bin, the color range is 16. In this phase, the histogram is used for changing the weight by the following formula.

$$\omega' = \frac{\omega}{1 + \sum_{j=1}^{3} \max[\ 0, H_{s,j}(b_j(e_{u,i,v})) - H_{e,j}(b_j(e_{u,i,v}))]} \ , \tag{4}$$

where H_s and H_e represent the histograms of input and output textures, respectively, $H(b)$ represents the amount of b color bin, and $b(e_{u,i,v})$ represents the interval that $e_{u,i,v}$ belongs to.

2.2 Model Cutting

In this subsection, it includes the cutting method of the cut-plane. First, the cut-plane equations are shown as follows.

$$f_1(x,y,z) = \begin{cases} y - y_0 = m_1(x - x_0), \\ z = z, \end{cases} \tag{5}$$

$$f_2(x,y,z) = \begin{cases} x - x_0 = m_2(z - z_0), \\ y = y, \end{cases} \tag{6}$$

$$f_3(x,y,z) = \begin{cases} z - z_0 = m_3(y - y_0), \\ x = x, \end{cases} \tag{7}$$

The equation (5)-(7) are plane equations for easily processing the problems of rotation and displacement. After obtaining cut-planes, to find the cells that intersect between the model and the plane equation, the equations (5)-(7) should be rewritten as follows.

$$f_1(x,y,z) = y - y_0 - m(x - x_0), \tag{8}$$

$$f_2(x,y,z) = x - x_0 - m(z - z_0), \tag{9}$$

$$f_3(x,y,z) = z - z_0 - m(y - y_0), \tag{10}$$

The results $f_k(v_{i1})$, $f_k(v_{i2})$, $f_k(v_{i3})$ can be categorized into three cases as formula (11).

$$\begin{cases} f_k(v_{i1}),\ f_k(v_{i2}),\ f_k(v_{i3}) > 0, \\ f_k(v_{i1}),\ f_k(v_{i2}),\ f_k(v_{i3}) < 0, \\ otherwise\ , \end{cases} \tag{11}$$

The first case corresponds to the block with dark blue color; the second case corresponds to the light blue color; the third case corresponds to the red color. The example is shown as Figure 1.

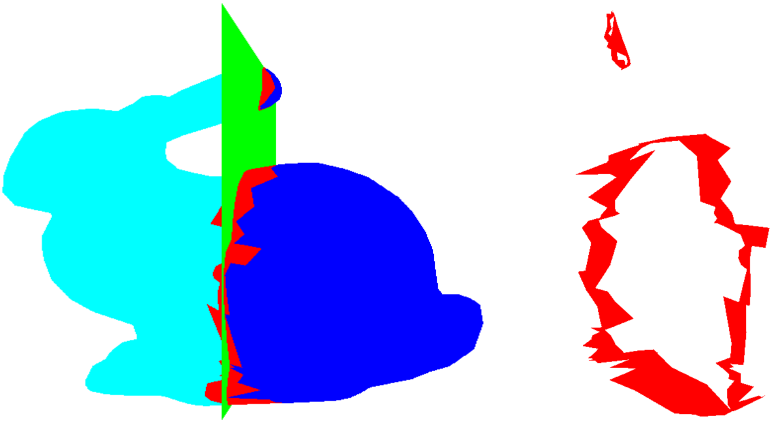

(a) (b)

Fig. 1. The graph of model cutting (a) The cut plane with green color, the cutting part with light blue color, the dark blue part without change, and the red part of several triangles. (b) The intersection parts.

3 Experimental Results

The experiment of the proposed method is stated in this section. The input image is with size 128×128, shown as Figure 2, and the output solid texture is with size 128×128×128, shown as Figure 3.

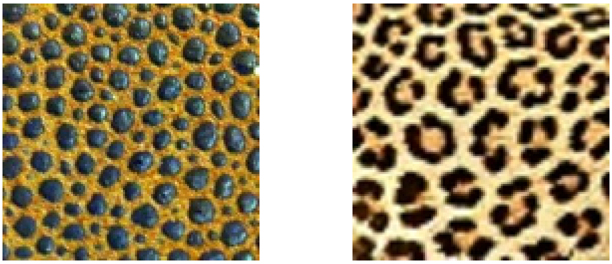
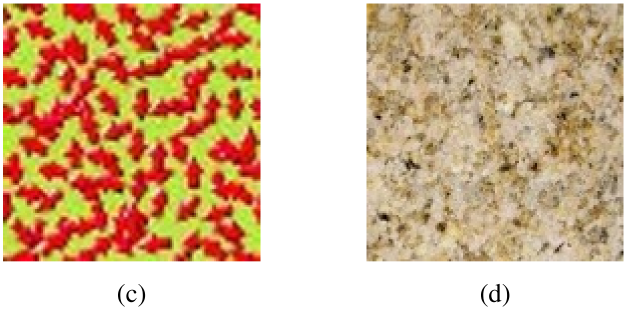
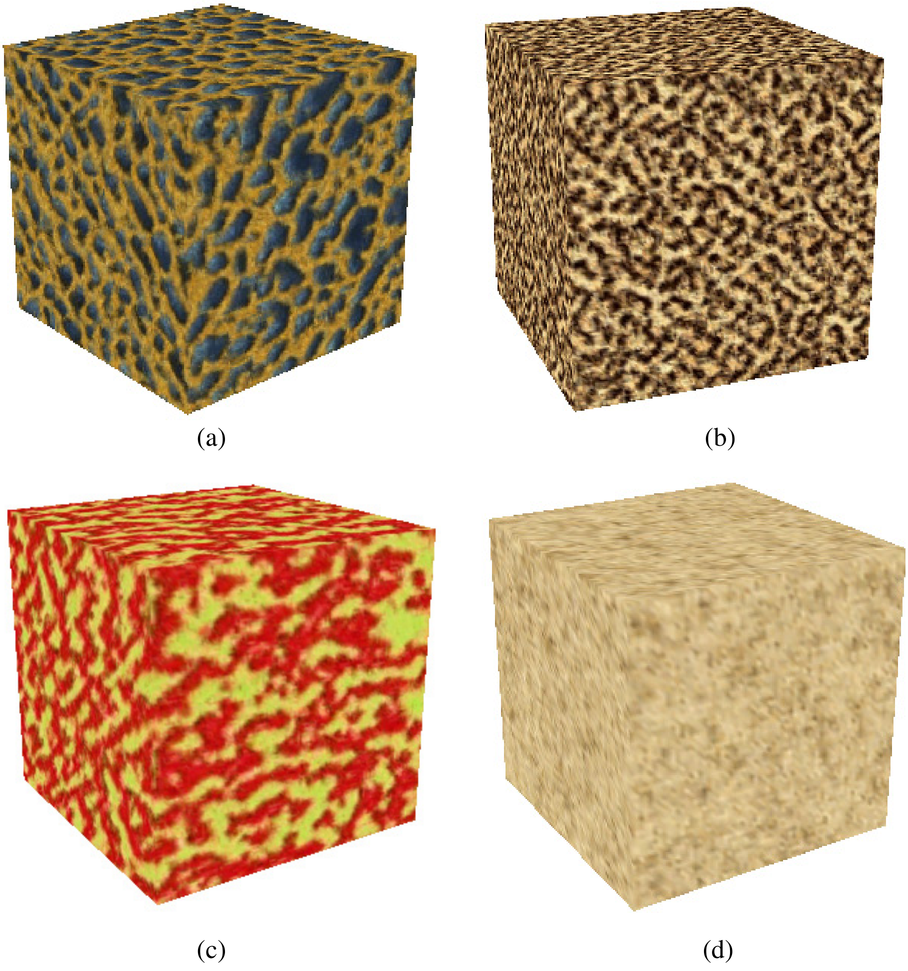

(a) (b)

Fig. 2. The input texture adopted in this paper



(c) (d)

Fig. 2. (*continued*)

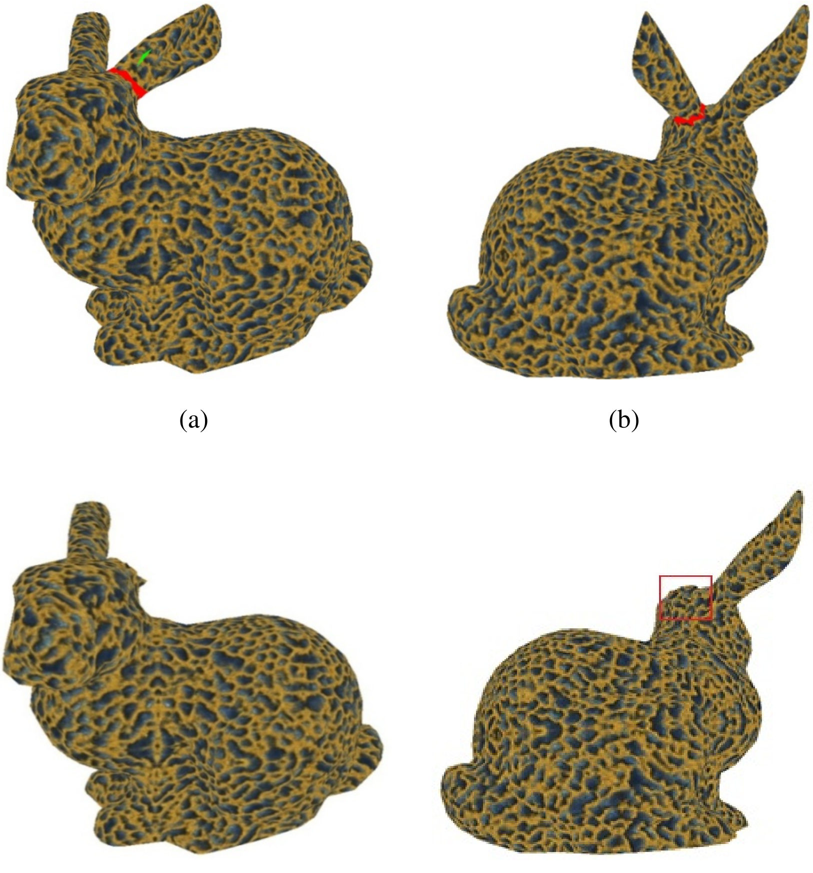


(a) (b)

(c) (d)

Fig. 3. The output solid texture

Another experiment is model cutting. The solid texture is shown as Figure 3(a). In Figure 3(a), the red part is selected by user and the green part is to abandon. Figure 3(b) is the back of Figure 3(a). Figure 3(c) and (d) are the front and the back of the solid texture of ear cut, respectively.



(a) (b)

(c) (d)

To test the speed of cutting, the adopted model are Zebra(20,157, 40,310)、Buddha(15,138, 30,272), Bunny1(5,114, 10,220), and Bunny2(34,817, 69,630). The two numbers are “numbers of vertices” and “numbers of cells”. The results are shown in Table 1. From Table 1, it is clear to see that the cutting time is directly proportional to the number of the cells. Overall, that is not time-consuming.

Table 1. The executing time of cutting

Zebra model		Buddha model	
The number of cells through the cut plane	time(ms)	The number of cells through the cut plane	time(ms)
78	0.6	38	0.4
138	0.7	140	0.6
397	0.9	358	0.7
662	1.2	392	0.8
798	1.7	527	0.9
1064	2	714	1.1
1147	2.3		

bunny1 model		bunny2 model	
The number of cells through the cut plane	time(ms)	The number of cells through the cut plane	time(ms)
65	0.2	333	1.4
112	0.3	405	1.7
236	0.4	576	1.8
339	0.6	681	2.3
		814	2.4

4 Conclusions

In this paper, we propose a method for mapping 2D texture to 3D solid mesh model. The proposed method uses the inner color information of synthesis solid texture to quickly fill the cut texture of the model. The proposed cutting algorithm uses plane equation to define cut plane. The advantage of this method is with the efficient processing time. When the mesh is very huge, the processing time of once cutting can be under one second and the seam looks nature. Therefore, the proposed method is very suitable for real time 3D rendering.

References

1. Perlin, K.: An image synthesizer. SIGGRAPH Computer Graphics 19, 287–296 (1985)
2. Kopf, J., Fu, C.W., Cohen-Or, D., Deussen, O., Lischinski, D., Wong, T.T.: Solid texture synthesis from 2D exemplars. ACM Transactions on Graphics 26(2), 2:1–2:9 (2007)
3. Dong, Y., Lefebvre, S., Tong, X., Drettakis, G.: Lazy solid texture synthesis. Computer Graphics Forum 27, 1165–1174 (2008)
4. Qin, X., Yang, Y.H.: Aura 3D textures. IEEE Transactions on Visualization and Computer Graphics 13, 379–389 (2007)
5. Du, S.P., Hu, S.M., Martin, R.R.: Semiregularsolid texturing from 2D image exemplars. IEEE Transactions on Visualization and Computer Graphics 19, 460–469 (2013)
6. Takayama, K., Okabe, M., Ijiri, T., Igarashi, T.: Lapped solid textures: filling a model with anisotropic textures. ACM Transactions on Graphics 27, 1–9 (2008)

Greedy Active Contour Detection for Median Nerve on Strain Sonographic Images

Chii-Jen Chen[1], You-Wei Wang[2], Sheng-Fang Huang[3], and Yi-Shiung Horng[4,5]

[1] Department of Computer Science and Information Engineering,
Yuanpei University, Hsinchu, Taiwan
[2] Department of Computer Science and Information Engineering,
National Taiwan University, Taipei, Taiwan
[3] Department of Medical Informatics,
Tzu Chi University, Hualien, Taiwan
[4] Department of Physical Medicine and Rehabilitation,
Taipei Tzu Chi Hospital, Buddhist Tzu Chi Medical Foundation, Taipei, Taiwan
[5] Department of Medicine,
Tzu Chi University, Hualien, Taiwan
cjchen@mail.ypu.edu.tw

Abstract. Carpal tunnel syndrome (CTS) is commonly occurred in occupations using vibrating manual tools or handling tasks with highly repetitive and forceful manual exertion. Recently, the ultrasonography has been used to evaluate CTS by monitoring median nerve movements. In order to facilitate the automatic extraction of shape characteristics for the median nerve, this paper designed a procedure that used greedy active contour detection model (GACD) to detect the edge of median nerve in ultrasound image. We selected a ROI to be an initial of virtual contour for median nerve in original ultrasound image. That can enhance the sensitivity of proposed GACD model to detect the contour of median nerve. In the experiment, the results show that the performance of the method is feasible and accurate.

Keywords: ultrasound, carpal tunnel syndrome (CTS), median nerve, greedy, active contour.

1 Introduction

Carpal tunnel syndrome (CTS) is a clinical disorder caused by compression of the median nerve at the wrist, which is commonly occurred in occupations using vibrating manual tools or handling tasks with highly repetitive and forceful manual exertion. The diagnosis of carpal tunnel syndrome (CTS) can rely on a combination of characteristic symptoms and electrophysiologic abnormalities. Nevertheless, an electrodiagnostic study remains an expensive and time-consuming procedure not readily accessible to many physicians who are encountering the disease.

In recent years, ultrasound imaging plays an important role in the diagnosis of CTS, because of its wide availability, lower cost, non-invasiveness, and shorter

H. Sun et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 329, DOI: 10.1007/978-3-319-12286-1_32

examination time [1]. Ultrasound has been shown to have a sensitivity as high as 94% and a specificity as high as 98% in the diagnosis of CTS, and can provide structural abnormalities and diagnostic reference in imaging [2-4], to make up for the lack of nerve electrical inspection. Many scholars have been trying to establish the ultrasound diagnostic criteria for the diagnosis of carpal tunnel syndrome and its use, including the measurement of median nerve cross-sectional area, flattening ratio, swelling ratio and palmar bowing of the flexor retinaculum, etc [5]. Dlley et al. used the cross-correlation between the images based on the wrists, elbows, shoulders and neck stretches, to measure the sliding elastic characteristics of the median nerve [6]. On the other hand, Yoshii et al. estimated the cross-sectional area of the median nerve, block aspect ratio, circularity, block perimeter and other characteristics, and then judge the differences and effectiveness of these features [7].

In order to facilitate the automatic extraction of shape characteristics for the median nerve, this paper presents a greedy active contour detection procedure, named GACD, to detect the contour of median nerve on strain sonographic images. We first chose a ROI to be an initial of virtual contour about median nerve in ultrasound image. This pre-processing can enhance the sensitivity of region contour, and assist to detect the contour of median nerve by proposed GACD model. In the experiment, the results show that the performance of the method is feasible and accurate.

2 The Proposed Contour Detection Model

In this section, we designed a procedure that used greedy active contour detection model (GACD) to detect the edge of median nerve in ultrasound image. The ultrasound image may have to do pre-processing to reduce noise and enhance contrast of the target region that region we called the region of interest (ROI). We used ROI to be an initial of virtual contour about median nerve in ultrasound image. The pre-processing can enhance the sensitivity of GACD to detect the contour of median nerve. Through the mechanism of convergence, we can obtain the contour of median nerve, the system framework such as Fig. 1. The purpose of system framework is able to make GACD obtain the contour of median nerve in ultrasound image. Because of ultrasound image had high noise and low contrast that were difficult to obtain the complete region of median nerve, so we used the GACD that always can segment the closed contour in ROI.

2.1 Data Acquisition

In this paper, there were 12 testing data which were supported by Department of Physical Medicine and Rehabilitation, Taipei Tzu Chi Hospital, Buddhist Tzu Chi Medical Foundation, Taipei, Taiwan. In each case, there were 220 continuous imaging slices and the scanning time was 20 seconds per case. The size of each imaging slice is 352 × 434 pixels. During the scanning procedure, six wrist motions, which include rest, straight, hook, fist, tabletop, and straight fist, must be completed by each patient within the time, as shown in Fig. 2. Then, the median nerve will be tightened and relaxed by the six wrist motions at different time points, and displayed in the continuous imaging slices. After the data acquisition, the contour of median nerve can be obtained by the proposed curve matching algorithm from the continuous imaging slices.

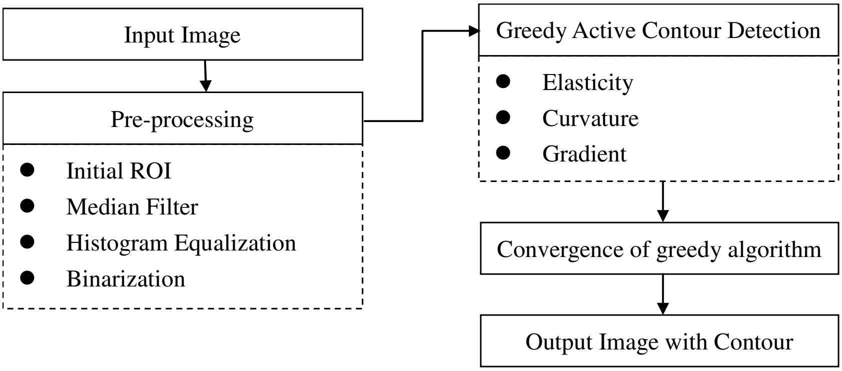


Fig. 1. The system framework of median nerve detection

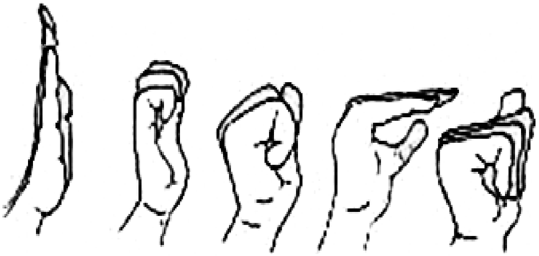


Fig. 2. The first five wrist motions are straight, hook, fist, tabletop, and straight fist. The final one, straight fist, is a motion simply for taking a rest break.

2.2 Image Pre-processing

Region of interest is called ROI that is a rectangle region. At the beginning in the procedure of our purpose, we used ROI to select the median nerve that was initial contour. The initial contour can help us to use GACD for contour segmentation. In GACD algorithm, we should have control points that are initial contour to assist GACD algorithm shown as Fig. 3. After determining ROI, ROI should have some pre-processing that could help the GACD algorithm be more accurate. First, because of the ultrasound images had high noise, so we used median filter to remove the pepper/salt noise in the ROI. Next, the ultrasound images always have low contrast that used histogram equalization to assist low contrast problem. Histogram equalization can enhance the contrast that means the ambiguous of ROI can be a high contrast image. Finally, we can accord the purpose which requires the ROI process to a binary image, so the gradient of ROI can be more easily to calculate. In this section, we introduce the main purpose of pre-processing that defines the process in the ROI to initial control points and some image processing.

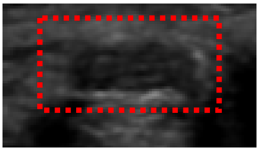


Fig. 3. The initial contour of ROI

2.3 Greedy Active Contour Detection

The proposed contour detection method, GACD, could be divided into three parts: elasticity (Ela), curvature (Cur), Gradient (Gra) [8]. These three criteria can assist to the GACD algorithm. The contour detection will be able to define the target contour that formula as following [8]:

$$GACD = \arg MIN(Ela + Cur + Gra) \tag{1}$$

where *Ela* is to decide the distance between control points that direct the movement of ROI. *Cur* is able to regulate the degree of bending. *Gra* means the gradient of image that can find out the most likely contour points as GACD. We used three criteria to decide ROI contour shrink iteration. The formula is searching at each control points around the mask (3*3). We choose the argument of minimum GACD be the control points of next iteration, then after several times iteration of GACD that contour will shrink on median nerve that is main purpose. In other words, the control points moved with mask (3*3) in each iteration to decide the movement such as Fig. 4. After each iteration, we examine the control points which will be too close, the two closed control points were removed and created the new control point at the center point between the two closed control point; If the distance of control point is too far from the adjacent control points, we will add the center point between the two neighbor adjacent control points that can reduce the program to repeat the calculate of control points.

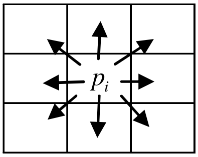


Fig. 4. The movement of control point

Elasticity is used adjacent control points to decide elasticity, calculated as following:

$$\bar{d} = \frac{\|p_N - p_1\| + \sum_{i=1}^{N-1} \|p_{i+1} - p_i\|}{N} \quad (2)$$

$$Ela = \left|\bar{d} - \|p_i - p_{i-1}\|\right| \quad (3)$$

where $\bar{d}$ is the average distance of each two control points. *Ela* is use the distance between control points to decide the best location of movement choice [9]. In other words, The Ela can make the distance between each control point to be average.

Curvature is used adjacent control points to regulate the degree of bending, calculated as following:

$$Cur = \|(p_{i+1} - p_i) + (p_{i-1} - p_i)\|^2 \quad (4)$$

where *Cur* is used both of the target control point and two neighbor control points to decide the bending condition. Fig. 5 is a schematic of curvature. We want the bend limit as well as smooth boundary, so the value of *Cur* required smaller that is the best [10].

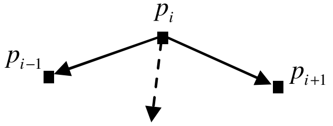


Fig. 5. A schematic of curvature

Gradient is used the edge detection of image processing that called the method of differential. We used the differential method to find the maximum of gradient that can be contour evidence, calculated as following:

$$g(x,y) = \sqrt{(I(x+1,y) - I(x,y))^2 + (I(x,y+1) - I(x,y))^2} \quad (5)$$

And normalized the gradient such as:

$$Gra = \frac{\max(g) - g(p_i)}{\max(g) - \min(g)} \quad (6)$$

where *Gra* is the normalized of gradient, max(*g*) is the maximum gradient in ROI, min(*g*) is the maximum gradient in ROI. Through the normalization, when the value of *Gra* is become small that means the control points is the contour.

2.4 Convergence of Greedy Algorithm

Finally, GACD algorithm required a stop condition. When the algorithm achieves convergence, then the edge detection was able to stop the iteration. The main purpose of procedure is to detect median nerve in ultrasound image. Median nerve is a closed curve and a similar internal texture. When the contour does not change in any iteration that means GACD is find out the median nerve contour shown in Fig. 6.

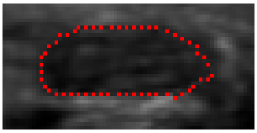


Fig. 6. The convergence of GACD

3 Experiment Result

In the experiments result, we have tested 12 median nerve cases in ultrasound image, and there are more than 220 images in each case. We proposed a procedure that could detect the contour of median nerve shown in Fig. 7. We prepared the initial control points on ROI that can move the control points where the control points become a median nerve contour. According to *Ela*, *Cur*, *Gra*, the movement of control points could change to convergence in GACD procedure. After the GACD procedure, the result was shown in Fig. 8 where red contour is median nerve contour. The results show that GACD procedure do image segmentation of median nerve that contour was purely.

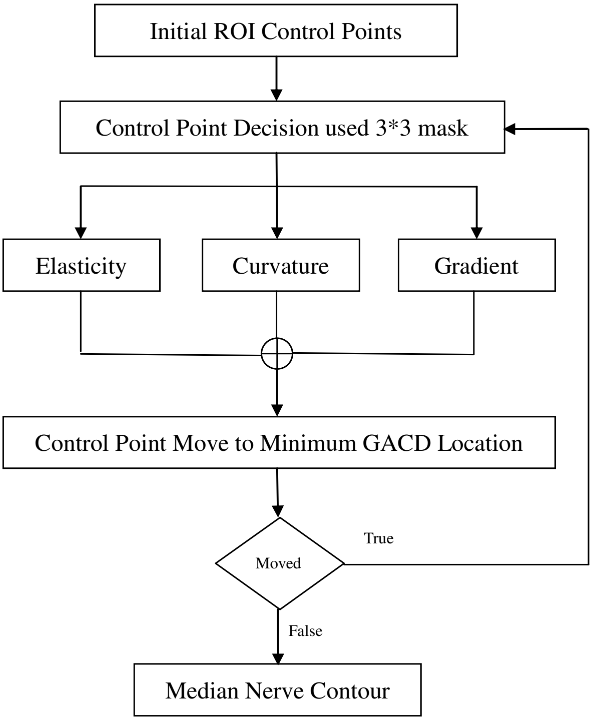

Initial ROI Control Points
Control Point Decision used 3*3 mask
Elasticity
Curvature
Gradient
Control Point Move to Minimum GACD Location
Moved
True
False
Median Nerve Contour

Fig. 7. The detail of GACD procedure

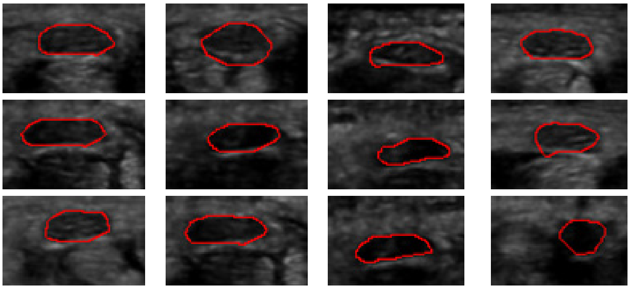


Fig. 8. The results of median nerve contour by proposed GACD procedure

4 Discussion and Conclusion

This paper presents a greedy active contour detection procedure, named GACD, to detect the contour of median nerve on strain sonographic images. We first chose a ROI to be an initial of virtual contour about median nerve in original image. This preprocessing can enhance the sensitivity of region contour, and assist to detect the contour of median nerve by proposed GACD model. The experimental results also show that the performance of the method is feasible and accurate.

However, there are limitations in the proposed GACD model. The computation of contours is still sensitive to the reference contour. The faulty contour in the reference image may lead to erroneous segmentation result. Therefore, we will improve the method of reference image selection by incorporating with more features in the future. Furthermore, we also need a more robust scheme to correct such erroneous propagation to improve the performance of contour tracking, and provide valuable structural information for the diagnosis of CTS.

References

1. Wong, S.M., Griffith, J.F., Hui Andrew, C.F., Lo, S.K., Fu, M., Wong, K.S.: Carpal Tunnel Syndrome: Diagnostic Usefulness of Sonography. Radiology 232, 93–99 (2004), doi:10.1148/radiol.2321030071; Published online before print May 20, 2004
2. Britz, G.W., Haynor, D.R., Kuntz, C., Goodkin, R., Gitter, A., Kliot, M.: Carpal tunnel syndrome: correlation of magnetic resonance imaging, clinical, electrodiagnostic, and intraoperative findings. Neurosurgery 37(6), 1097–1103 (1995)
3. Duncan, I., Sullivan, P., Lomas, F.: Sonography in the diagnosis of carpal tunnel syndrome. AJR Am. J. Roentgenol. 173(3), 681–684 (1999)
4. John, V., Nau, H.E., Nahser, H.C., Reinhardt, V., Venjakob, K.: CT of carpal tunnel syndrome. AJNR Am. J. Neuroradiol. 4(3), 770–772 (1983)
5. Keles, I., Karagulle Kendi, A.T., Aydin, G., Zog, S.G., Orkun, S.: Diagnostic precision of ultrasonography in patients with carpal tunnel syndrome. Am. J. Phys. Med. Rehabil. 84(6), 443–450 (2005)

6. Dilley, A., Lynn, B., Greening, J., DeLeon, N.: Quantitative in vivo studies of median nerve sliding in response to wrist, elbow, shoulder and neck movements. Clin. Biomech. 18(10), 899–907 (2003)
7. Yoshii, Y., Villarraga, H.R., Henderson, J., Zhao, C., An, K.N., Amadio, P.C.: Ultrasound assessment of the displacement and deformation of the median nerve in the human carpal tunnel with active finger motion. J. Bone Joint Surg. Am. 91(12), 2922–2930 (2009)
8. Tiilikainen, N.P.: A Comparative Study of Active Contour Snakes, pp. 21–26. Copenhagen University, Denmark (2007)
9. Castleman, K.R., Riopka, T.P., Qiang, W.: FISH image analysis. IEEE Engineering in Medicine and Biology Magazine 15, 67–75 (1996)
10. Williams, D.J., Shah, M.: A fast algorithm for active contours and curvature estimation. CVGIP: Image Underst. 55, 14–26 (1992)

One Decomposition Method of Concept Lattice

Haixia Li[1], Jian Ding[1], Dongming Nie[1], and Linlin Tang[2]

[1] Department of Public Teaching, Anhui Xinhua University, Hefei 230088, China
learain1@126.com
[2] Harbin Institute of Technology Shenzhen Graduate School, Shenzhen 518055, China

Abstract. A method of decomposition into sub-concept lattices with the same attributes set is proposed. Based on the decomposition function, it is proved that when all the nodes of the concept lattice are arranged in ascending order of the extent base, the node of the sub-concept lattice is generated and its intent is equal to the intent of the node decomposed if the extent of the first decomposition; at the same time, the Hasse diagram of the sub-lattices is generated according to the route number of the node generated, then the method is described. Finally, the effectiveness of the developed method is illustrated by an example.

Keywords: concept lattice, sub-concept lattices, Hasse diagram, extent, intent, decomposition function.

1 Introduction

Concept lattice[1] and formal concept analysis is put forward in 1982 by Professor Wille of Germany. Since the concept lattice constructed by the formal context does not artificially reduce the complexity and contains all the details of the data, it is an effective knowledge representation and knowledge discovery tool, and more and more attention has been paid to its inherent advantage. In recent years, concept lattice has made considerable progress in software engineering, knowledge discovery, information retrieval and many other fields[2-6].

In real life, sometimes dealing with the data of huge quantity and complex relationships need cost a lot of manpower, material resources, time, etc., and sometimes, just selecting the data of interest is okay. Therefore, if the concept lattice generated by the data background is decomposed into several sublattices, dealt with separately, it easy to solve data analysis and rule extraction in case of huge data, save manpower and material resources and improve processing efficiency. The decomposition of concept lattice can be applied to data mining, software engineering, software re-structure and other fields[7,8]. However, up to now, domestic and foreign scholars have not fully studied it. In this paper, based on the decomposition function, one method of decomposition into sublattices with the same attributes set is discussed, this method not only eliminates the redundant nodes and reduces the number of comparisons, but also greatly improve the decomposition efficiency of concept lattice.

H. Sun et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 329, DOI: 10.1007/978-3-319-12286-1_33

2 Preliminary

Definition 1 [7]. One formal context $K = (O, D, R)$ is a triple, O is an object set, D is an attribute set, $R \subseteq O \times D$ is binary relation between O and D . For $A \subseteq O$, $B \subseteq D$,the mapping is defined as

$$A' = \{m \in D \mid \forall g \in A, (g, m) \in R\} ;$$
$$B' = \{g \in O \mid \forall m \in D, (g, m) \in R\} .$$

If $A = B', B = A'$, (A, B) is called one node, and A is the extent of the node (marked as extent(C)), B is the intent(marked as intent(C)).

Definition 2 [7]. If $C_1 = (A_1, B_1)$ and $C_2 = (A_2, B_2)$ is two nodes and $A_1 \supseteq A_2$ ($\Leftrightarrow B_1 \subseteq B_2$), C_2 is called the child node of C_1, C_1 is the father node of C_2 ,which is denoted as $C_1 \geq C_2$. If there is no node C_3 which satisfies $C_1 \geq C_3 \geq C_2$, C_2 is called the direct child node of C_1, C_1 is the direct father node of C_2 . According to this sequence, the set composed is called the concept lattice of K and denoted as $L(O, D, R)$.

For example, corresponding to the formal context 1 in Table 1, the Hasse diagram of the concept lattice is shown in Figure 1, where the digital expresses objects, the letter expresses attribute, “1" denotes relationship between the two, “0” denotes no.

Table 1. Formal context 1

	a	b	c	d	e
1	1	0	1	1	0
2	1	1	0	1	0
3	0	1	1	0	0
4	0	1	1	0	1
5	1	0	1	1	1

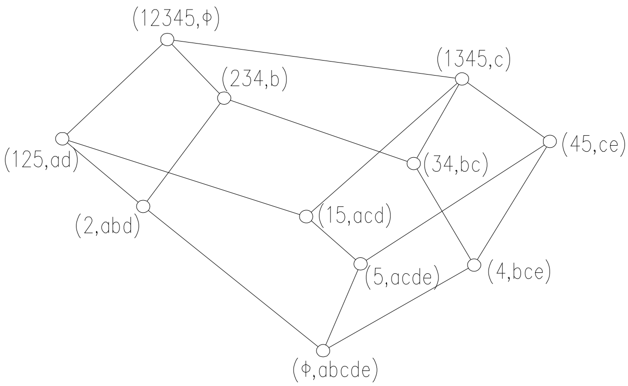


Fig. 1. Hasse diagram of the formal context 1

Theorem 1 [7]. For the concept lattices $L(K_1)$ and $L(K_2)$, where $K=(O,D,R)$, $K_1=(O_1,D,R_1)$, $K_2=(O_2,D,R_2)$, $O_1\cup O_2=O$, $O_1\cap O_2=\phi$, then $L(K)$ can be mapped to sublattice $L(K_1)$ and $L(K_2)$ through the function φ :

$$\phi((A,B))=\left(\left(A\cap O_1,(A\cap O_1)'\right),\left(A\cap O_2,(A\cap O_2)' A\cap O_2,(A\cap O_2)'\right)\right) \tag{1}$$

3 The Decomposition Method of Concept Lattice

3.1 Basic Definitions and Theorems

Definition 3. For the concept lattices $L(K_1)$ and $L(K_2)$, the function φ above is called decomposition function of concept lattice $L(K)$.

Theorem 2. For $\phi(C)=(C_1,C_2)$, $C\in L(K)$, $C_i\in L(K_i), i=1,2$, if the extent of the nodes generated by all child nodes of the node C is not equal to the extent(C_i)($(\mathrm{i}=1\mathrm{or}2)$), then intent$(\mathrm{C_i})$=intent(C).

Proof. Let $\mathrm{C}=(\mathrm{A}, \mathrm{B})$, by Equation (1),

$$\phi((A,B))=((A_1,A'),(A_2,A_2')),$$

it is clear that $A=(A\cap O_1)\cup(A\cap O_2) = A_1\cup A_2$, $B\subseteq A_i', i=1,2$. If $A_1'\neq B$, that is, $B\subseteq A_1'$, according to the overlay of context [7], (A_1,A') and (A_2,A_2') will generate one node the intent of which is greater than B , Suppose that the node is $C_0=(A_0,B_0)$, then $C_0\in L(K)$ and C_0 is the child node of C , therefore $A_0\subset A$ and $A_1=A\cap O_1\supseteq A_0\cap O_1\supseteq A_1\cap O_1=A_1$ hold, so it follows that $A_0\cap O_1=A_1$, that is, decomposition of C_0 generates a node of which extent is A_1 , which contradicts the known condition , thus $A_1'=B$, that is, intent$(C_1)=B$.

In the same way, if the extent of the nodes decomposed by all child nodes of the node C is lesser than extent(C_2),we can also conclude that intent$(C_2)=B$.

Theorem 3. All the nodes of the concept lattice are arranged in ascending order of the extent base, one concept lattice is decomposed into sub-concept lattices with the same attributes sets. If the extent of the node C_1 generated from the decomposition of the node C already exists, then the node C_1 can be ignored in the process of the decomposition of C .

Proof. Since all the nodes of the concept lattice are arranged in ascending order of the extent base, if the extent of the node C_1 generated from the decomposition of the node C already exists, which is denoted as A , then the extent A must be generated from the decomposition of the child node of the node C (let the node C_0) or the child node (let C')of both C_0 and C .Taking Theorem 2 into account, we can deduce that A′=intent(C_0) or A′=intent(C'), that is, the node generated in sublattice is $(\mathrm{A}, \mathrm{intent}(C_0))$ or $(\mathrm{A}, \mathrm{intent}(C'))$ instead of $(\mathrm{A}, \mathrm{intent}(C))$. Therefore, the node C_1 can be ignored in the process of the decomposition of the node C . The proof is complete.

Theorem 4. One concept lattice is decomposed into sub-concept lattices with the same attributes sets if its nodes are arranged in ascending order of the extent base, then the child nodes of one node of sublattices are absolutely generated before its father nodes.

Proof. For any node $C_1=(A_1,B_1)\in L(K)$ and $\phi(C_1)=(C_{11},C_{12})$, where $C_{11}\in L(K_1)$, $C_{12}\in L(K_2)$, let C_{21} is any child node of C_{11}, in the following, we will prove that C_{21} is generated before C_{11}.

If C_{21} is generated by the child node of C_1, it is easy to see that C_{21} is generated before C_{11}; if C_{21} is generated by another node C_2 which is marked as (A_2,B_2) and is not child node of C_1, by definition 3, extent$(C_{21})=A_2\cap O_1$, and on the one hand extent (C_{21}) $\subset$ extent (C_{11}) = $A_1\cap O_1$, then extent (C_{21}) =extent (C_{11}) $\cap$ extent (C_{21}) = $(A_1\cap O_1)\cap(A_2\cap O_1)$ = $A_1\cap A_2\cap O_1$;on the other hand $\wedge(C_1,C_2)=\left(A_1\cap A_2,(A_1\cap A_2)'\right)$ ($\wedge(C_1,C_2)$ is the child node of both C_1 and C_2), we can get $\left(\left(A_1\cap A_2\cap O_1,(A_1\cap A_2\cap O_1)'\right),\left(A_1\cap A_2\cap O_2,(A_1\cap A_2\cap O_2)'\right)\right)$, therefore C_{21} can be generated from $\wedge(C_1,C_2)$ and it is already generated before C_{11}.This completes the proof.

There are two major problems during constructing concept lattice: generating nodes and edge generation. If one concept lattice is decomposed into sub-concept lattices with the same attributes sets, just its nodes are arranged in ascending order of the extent base, and it is follows from theorem 2 and 3 that the node of the sublattice is generated and its intent is equal to the intent of the node decomposed if the extent of the decomposition appears for the first time; otherwise it is ignored, this will solve the problem of generating nodes in sub-concept lattice; by theorem 4, child nodes of sublattices are absolutely generated before its father nodes, if the nodes in sublattice are numbered according to the generation sequence, then the route numbers are generated from whether the direct father-son relationship, finally the edges are connected by the route, so the Hasse diagram of the sub-lattices can be easily obtained.

3.2 Algorithm Description

In this section, we present the algorithm of one concept lattice $L(K)$ which is decomposed into sublattices with the same attributes sets, where $\|L(K)\|=m$. Its steps are as follows:

① Arrange and number all the nodes of $L(K)$ in ascending order of the extent, and let the variable $i=1$;

② Let $L(K_{\mathrm{i}}) = \varphi, L(K_i') = \{(\phi,t)\},\ \ L(K_i'') = \{\phi\},\ \ t = 0, j = 1$;

③ Remove a node (A_j, B_j) of $L(K)$,and mark $A_j \cap O_i$ as $\mathrm{ext}(A_j, O_i)$,if there exists one node of which extent is $\mathrm{ext}(A_j, O_i)$, let $j = j+1$, keep the step;

④ Otherwise let $t = t+1$, $(A,B) = (A_j \cap O_i, B_j)$, store the node in $L(K_i)$ and $(A, B, (t))$ in $L(K_i')$;

⑤ Find all direct child nodes of $(A, B, (t))$ in $L(K_i')$, add number (t) after any one route of each direct child node as the route to that node. Store the nodes with route in $L(K_i'')$. If $j = m$, let $i = i+1$, turn to the step ②; Otherwise let $j = j+1$, turn to the step ③;

⑥ Until $i = 2$, take out all elements of $L(K_i'')$ respectively, connect edges according to the path, the Hasse diagram of the sub-lattices can be obtained, and all nodes of sub-lattices in $L(K_i)$

4 Example

Decompose the concept lattice corresponding to context 1 into two sub-concept lattices with the same attribute set, where $O_1 = \{1,2,3\},\ \ O_2 = \{4,5\}$. The imagines of the nodes in the decomposition function φ are as follows, where the nodes underlined are the ones of sublattices which extent appears for the first time.

$$\phi(\varphi, abcde) = \left(\underline{(\varphi, abcde)}, \underline{(\varphi, abcde)}\right),$$

$$\varphi(5, acde) = \left(\left(\{5\} \cap O_1, (\{5\} \cap O_1)'\right), \left(\{5\} \cap O_2, (\{5\} \cap O_2)'\right)\right)$$

$$= \left((\phi, \mathrm{abcde}), \underline{(5, acde)}\right),$$

$$\varphi(4, bce) = \left(\left(\{4\} \cap O_1, (\{4\} \cap O_1)'\right), \left(\{4\} \cap O_2, (\{4\} \cap O_2)'\right)\right)$$

$$= \left((\phi, abcde), \underline{(4, \mathrm{bce})}\right),$$

$$\varphi(2, abd) = \left(\underline{(2, abd)}, (\phi, abcde)\right) ,$$

$$\varphi(15, acd) = \left(\left(\{15\} \cap O_1, (\{15\} \cap O_1)'\right), \left(\{15\} \cap O_2, (\{15\} \cap O_2)'\right)\right)$$

$$=\left(\underline{(1,\ \mathrm{acd})},(5,acde)\right),\ \text{where}\ \{5\}'\neq\{\mathrm{acd}\},$$

$$\phi(45,\ \mathrm{ce})=\left((\phi,abcde),\underline{(45,\ ce)}\right),$$

$$\phi(34,bc)=\left(\underline{(3,\ \mathrm{bc})},(4,bce)\right),$$

$$\phi(234,b)=\left(\underline{(23,\ \mathrm{b})},(4,bce)\right),$$

$$\phi(125,ad)=\left(\underline{(12,ad)},(5,acde)\right),$$

$$\phi(1345,c)=\left(\underline{(13,c)},(45,ce)\right),,$$

$$\phi(12345,\varphi)=\left(\underline{(123,\varphi)},(45,ce)\right).$$

The decomposition process can be illustrated by Table 2.

Table 2. Decomposition process

Node (A,B)	ext $(A\cap O$	Node generated	No.	Direct child node	route	Node with route
$(\varphi,abcde)$	ϕ	$(\varphi,abcde$	①	--	①	$(\phi,abcd$,{①})
$(5,acde)$	ϕ	no				
$(4,bce)$	ϕ	no				
$(2,abd)$	$\{2\}$	$(2,abd)$	②	①	①②	$(2,abd,${①②})
$(15,acd)$	$\{1\}$	$(1,acd)$	③	①	①③	$(1,acd$,{①③})
$(45,ce)$	ϕ	no				
$(34,bc)$	$\{3\}$	$(3,bc)$	④	①	①④	$(3,bc$,{①④})
$(234,b)$	$\{23\}$	$(23,b)$	⑤	②,④	①②⑤, ①④⑤	($23,b$, {①②⑤ , ①④⑤})
$(125,ad)$	$\{12\}$	$(12,ad)$	⑥	②, ③	①②⑥, ①③⑥	($12,ad$,{①②⑥, ①③⑥})
$(1345,c)$	$\{13\}$	$(13,c)$	⑦	③,④	①③⑦, ①④⑦	($13,c$, {①③⑦ , ①④⑦})
$(12345,\phi)$	$\{123\}$	$(123,\phi)$	⑧	⑤,⑥,⑦	①②⑤⑧, ①②⑥⑧, ①③⑦⑧,	($123,\phi$, {①②⑤⑧,①②⑥⑧,①③⑦⑧})

At the same time, the decomposed nodes in $L(K)$ correspond to the generated nodes with route in sublattice $L(K_2)$ as follows: $(\varphi, abcde)$ → ($\varphi, abcde$,{①}); $(5, acde)$ → $(5, acde,$ {①②}); $(4, bce)$ → $(4, bce,$ {①③}); $(2, abd)$, $(15, acd)$ →no ; $(45, ce)$ →($45, ce$,{①②④, ①③④}.

In this paper, based on the decomposition function, one method of decomposition into sublattices with the same attributes set is proposed. If all nodes of one concept lattice are arranged in ascending order of the extent base, then child nodes in sub-concept lattices are absolutely generated before their father nodes; only compared with the extent of the direct child nodes in sublattices, if the extent decomposed appears for the first time, the node is generated, otherwise is not generated, the comparison with the extent of all nodes generated before is unnecessary. This will not only eliminate the redundant nodes and reduces the number of comparisons, but also greatly improve the decomposition efficiency of concept lattice. In literature [7], there will be a lot of redundancy during the decomposition, which will greatly increase the complexity of the problem itself; in Literature [8], the decomposition algorithm proposed contains only an object or attribute, when the objects and attributes of interest are more, the decomposition of efficiency may be greatly reduced.

5 Conclusion

The problem of decomposition of concept lattice is discussed in this paper. A new method is proposed that concept lattice is decomposed into sub-concept lattices with the same attributes sets. An example indicates our method is efficient and feasible. For rule extraction after decomposing and improving the decomposition efficiency will be further study.

Acknowledgement. The authors thank for the support from the project named Research on Multiple Description Coding Frames with Watermarking Techniques in Wavelet Domain which belongs to the NSFC (National Natural Science Foundation of China) with the Grant number 61202456, thank for the support from Anhui Province Natural Science project with the Grant number KJ2013B107, Excellent course of Xinhua University with the Grant number 2012jpkcx03.

References

1. Wille, R.: Restructing lattice theory: an approach based on hierarchies of Concepts, pp. 445–470. Reidel Publishing Company, Dordrecht (1982)
2. Wang, D., Xie, Q., Huang, D., Yuan, H.: Analysis of association rule mining on quantitative concept lattice. In: Lei, J., Wang, F.L., Deng, H., Miao, D. (eds.) AICI 2012. LNCS (LNAI), vol. 7530, pp. 142–149. Springer, Heidelberg (2012)

3. Eklund, P., Villerd, J.: A survey of hybrid representations of concept lattices in conceptual knowledge processing. In: Kwuida, L., Sertkaya, B. (eds.) ICFCA 2010. LNCS, vol. 5986, pp. 296–311. Springer, Heidelberg (2010)
4. Yang, L., Xu, Y.: A decision method based on uncertainty reasoning of linguistic truth-valued concept lattice. International Journal of General Systems 39(3), 235–253 (2010)
5. Stumme, G., Taouil, R., Bastide, Y., et al.: Computing iceberg concept lattices with TITANIC. Data & Knowledge Engineering 42, 189–222 (2002)
6. Abderrahim, E., Ennouary, Y.: Formal Concept Analysis for Information Retieval. International Journal of Computer Science and Information Security (IJCSIS) 7(2), 119–125 (2010)
7. Ganter, B., Wille, R.: Formal Concept Analysis. Springer (1999)
8. Bai, X., Ma, Y., Zhang, X.-P.: Concept lattice of simplest subdirect concept and algorithm. Journal of Anshan University of Science and Technology 27(6), 428–431 (2004) (in Chinese)

The Evaluation Method for Enterprise Group Information Planning and Implementation Based on PDCA Cycle

Weixiong Chen, Lailong Zou, Juan Liu, and Xiaochen Yang

Centre of Information Technology, CGN Power Co., Ltd., Shenzhen City, China
{chenweixiong,zoulailong,liujuan2013,yangxiaochen}@cgnpc.com.cn

Abstract. Lack of strategic information planning and implementation assessment methods, especially the effect and the completion rate for enterprise information planning, this paper puts forward an evaluation method called EIP (Planning and Implementation Evaluation) model based on PDCA Cycle, it carries out in planning, doing, checking and action. This paper also proposes a Comprehensive Evaluation Indicators (CEI) formula which can compute the completion rate in line with the information planning target. As preliminary results show the evaluation method can enhance the accuracy of the implementation planning as scheduled, and also play an important role in guiding the enterprise information implementation.

Keywords: Strategic Information Planning, PDCA Cycle, Comprehensive Evaluation Indicators.

1 Introduction

Nowadays the importance of strategic system planning is accepted by many large enterprise groups which have published the Twelfth Five-Year Strategic Information Plans. However, there are many problems in the information construction process, which has a big gap between the actual implementation and system planning, and there is a noticeable lack of refinement feasible implementation plan, and lack of regular reports for information planning evaluation which reports the effect and the completion rates of information implementation.

2 The Current Evaluation Method

There are some domestic and foreign researches in information systems strategic planning and evaluation. Albert proposed a second-order factor model (SISP) Evaluation model based on the conventional measurement methods and modern information technology statistics, it proved that the evaluation model has a strong operability [1]. Barry established a comprehensive information system organizational performance assessment framework which used the existing IS assessment theory, and combined

H. Sun et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 329, DOI: 10.1007/978-3-319-12286-1_34

with measurement concepts from other disciplines [2]. Bhanu proposed a model of information systems planning that can match the success of information systems [3]. Jerry proposed an information system performance evaluation methods based on Information Systems Functional Scorecard [4]. Xiao studied the comprehensive evaluation Indicators system and evaluation method of distribution network planning [5]. Ye also proposed assessment indicators for the construction of information system life cycle [6].

Throughout the above information planning and evaluation methods, they are some theoretical reference values for information planning and implementation, however, these methods cannot be scientifically assessed the effects of information technology planning, and could not reasonably assess the completion of the planning objectives, they need to carry out further study in operational aspects of information planning and implementation evaluation. Therefore, with years of research in information planning and evaluation methodology, we propose one enterprise groups evaluation method called EIP model based on PDCA Cycle.

3 Introduction of EIP Model

PDCA (Plan–Do–Check–Act) cycle is the process of quality management experts from the United States Deming, which is a quality plan development and organizational implementation [7]. It is not only used in business for the control and continuous improvement of processes and products, but also applied in all progressive management.

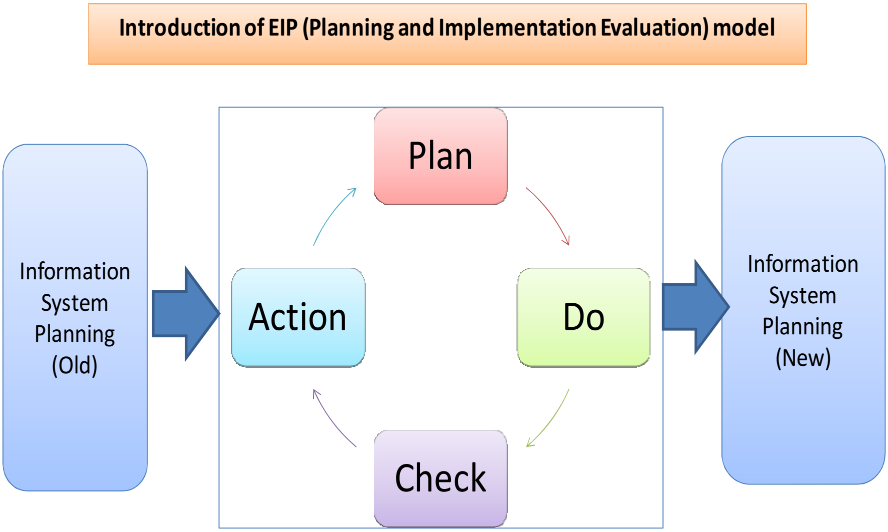

Introduction of EIP (Planning and Implementation Evaluation) model
Information System Planning (Old)
Plan
Do
Check
Action
Information System Planning (New)

Fig. 1. The Introduction of EIP Model

The basic idea of EIP evaluation model is the same with PDCA cycle, information planning and implementation of evaluation is actually a PDCA quality improvement closed-loop process. The EIP model is inputted from the Five-Year Plan of large enterprise group, which includes Information Plan (PLAN), Implementation of IT Plan (DO), the Check of IT plan (CHECK), the Action of IT plan (ACTION) four stages.

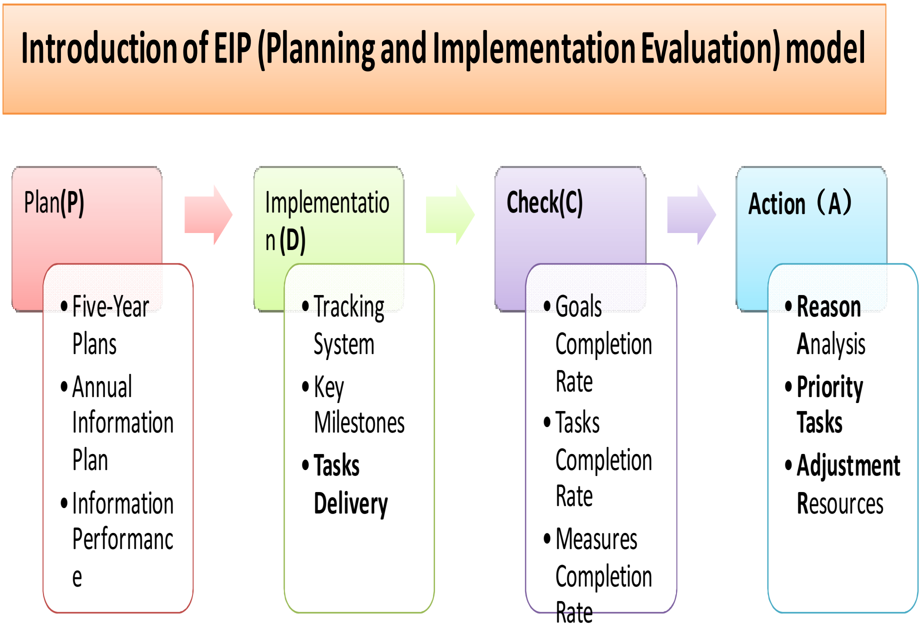


Fig. 2. The Detail Introduction of EIP Model

3.1 Annual Information Plan (PLAN)

The annual information plan undertakes the Twelfth Five-Year Strategic Information Plans of enterprise group, and it appoints clearly each responsibility units and their annual information performance. From the content point of view, the information plan tasks include information strategy goals, information strategic tasks, information strategic measures, information human and financial resources, information performance evaluation indicators etc.

Information strategic goals can be broken down a lot of information strategic tasks and information strategic measures, the information strategic tasks focuses on the specific task or project-specific information, while the information strategic measures refers to the specific management method to achieve information goals. Information strategic goals cannot be performed to achieve the goal itself, but it can be implemented through information strategic tasks or information strategic measures. Therefore, information strategic tasks or information strategic measures is important and an indispensable component of information strategic planning framework.

3.2 Implementation of Information Strategic Plan

In information implementation phase, we build our tracking system of information strategic planning, so as to monitor the implementation progress of information tasks and measures, and take review meeting to control the key milestones of them.

3.3 Checking of Information Strategic Plan

In the regular inspection and evaluation stage, we calculate the overall completion of information planning in real-time statistics, which includes information goals completion rate, information tasks completion rate and information measures completion rate.

Information goals completion rate is the actual percentage of completion information planning various sub-goals and combined score according to some weight percentage. Information tasks completion rate is the quantity ratio of information strategic tasks completed on schedule. Information measures completion rate is the ratio of the number of information major measures completed on schedule.

3.4 Corrective Action of Information Strategic Plan

When there are some milestones delayed or quality problems during the examination of information plan progress, we analyze the reason and take corrective action through putting more manpower and financial resources, submitting the senior decision-making meeting, strengthening the coordination of information, stopping information implementation and other means, so that the information plan can achieve as scheduled.

4 Comprehensive Evaluation Indicators

In order to assess the differences between information planning goals and the actual goals, we need to calculate the completion rate of information strategic planning goals. Studies have been shown that there is no mature information planning calculation method, we propose an information planning CEI Indicators (Comprehensive Evaluation Indicators) formula as followings:

$$\mathrm{CEI}(i)=\mathrm{SGC}(i)*\mathrm{W}_1+\mathrm{STC}(i)*\mathrm{W}_2+\mathrm{SIC}(i)*\mathrm{W}_3+\mathrm{KIC}(i)*\mathrm{W}_4 \tag{1}$$

The above formula is explained below for each parameter.

- SGC(i) refers to the Strategic Goal Completion Rate of enterprise group in the i year of the Twelfth Five-Year Strategic Information Plans, if the information strategic objectives completed ahead of schedule, the SGC can be considered 100%, here we can set its weight $W_1 = 15\%$;
- STC(i) refers to the Strategic Task Completion Rate of enterprise group in the i year, it mainly to assess the number ratio of strategic tasks completed on schedule, here we can set its weight $W_2 = 50\%$;
- SIC(i) refers to the Strategic Measures Completion Rate, it mainly to assess the number ratio of strategic measures completed on schedule, here we can set its weight $W_3 = 15\%$;
- KIC(i) refers to the Key Indicators Completion Rate of enterprise group in the i year, it includes information strategic goals, information strategic tasks, information strategic measures etc., here we can set its weight $W_4 = 20\%$;

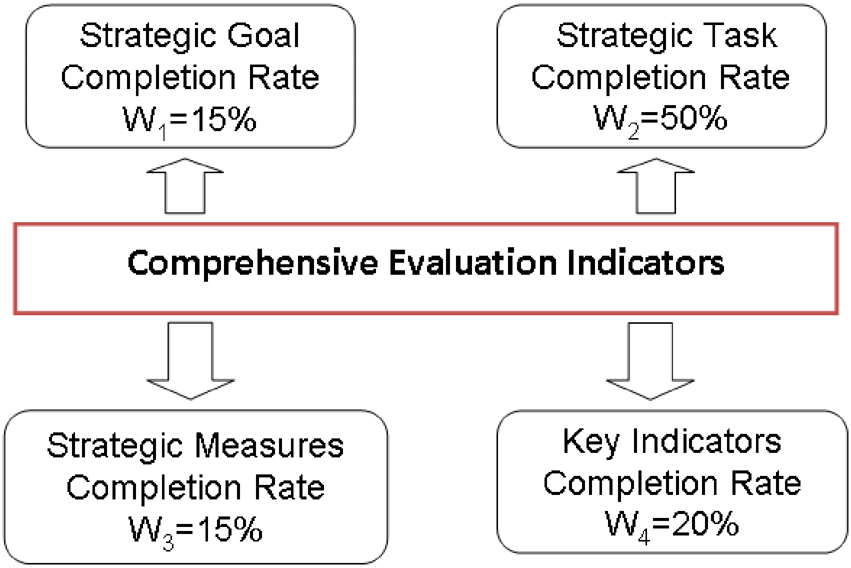

Strategic Goal Completion Rate W1=15%
Strategic Task Completion Rate W2=50%
Comprehensive Evaluation Indicators
Strategic Measures Completion Rate W3=15%
Key Indicators Completion Rate W4=20%

Fig. 3. Comprehensive Evaluation Indicators

5 Experimental Results

In accordance with the CEI evaluation formula, we have calculated the implementation completion rate of an enterprise information planning from the year of 2011 to 2013, which we consider its completion rate of strategic goals, strategic task, strategic measures, and key indicators, therefore, we predict the CEI in the year of 2014 and 2015, which can test the progress in the implementation of information technology planning, the annual planning goals is 95 in the year of 2015, the CEI predict value is 97.55, it shows that the strategic planning goals can be completed on schedule in the year 2015.

Table 1. The Comprehensive Evaluation Indicators

	Weight	2011	2012	2013	2014	2015
Target Value		75	80	85	90	95
Actual Value		72.25	73.55	81.85	89.25	97.55
Strategic Goal Completion rate	15%	65	69	84	86	99
Strategic Task Completion Rate	50%	70	73	83	90	98
Strategic Initiatives Completion Rate	15%	70	74	73	89	94
Key Indicators Completion Rate	20%	85	78	84	90	98
CEI Value		72.25	73.55	81.85	89.25	97.55

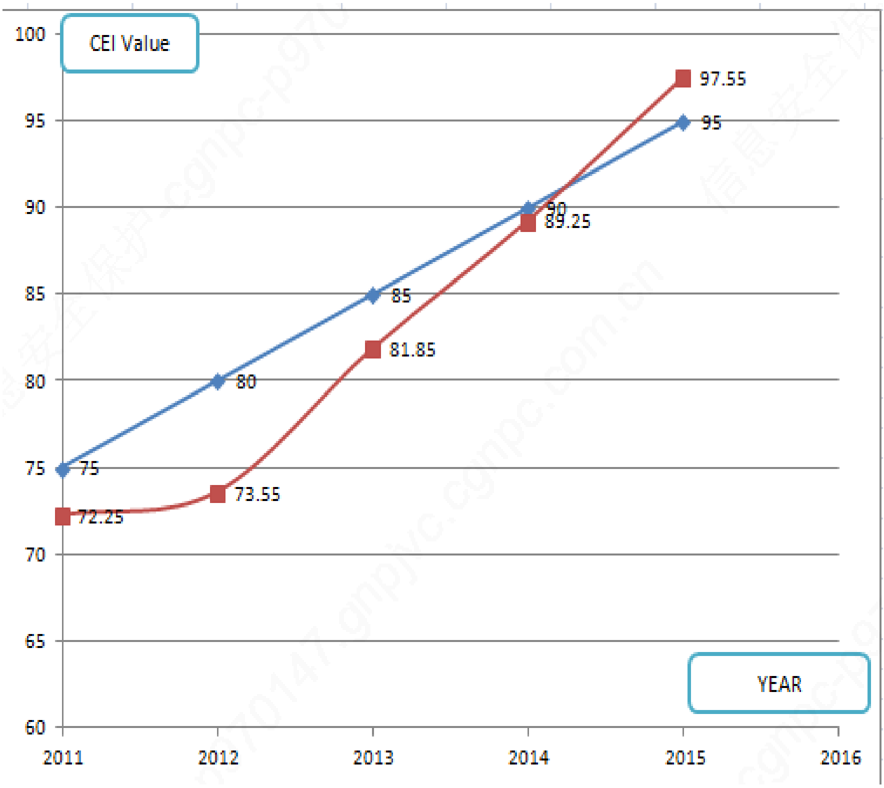


Fig. 4. The Experimental Results of Comprehensive Evaluation Indicators

6 Summary

The EIP evaluation method is constantly improving in concepts and methods, and it is used in a large enterprise group strategic planning projects. As preliminary results show the evaluation method can enhance the accuracy of the implementation direction planning as scheduled, and also play an important role in guiding the enterprise group's implementation.

References

1. Albert, H.S., Varun, G.: Strategic Information Systems Planning Success: An Investigation of the Construct and Its Measurement. MIS Quarterly 22(2), 139–163 (1998)
2. Barry, L.M., Leon, A.K., Victor, R.P.: A Comprehensive Model for Assessing the Quality and Productivity of the Information Systems Function: Toward a Theory for Information Systems Assessment. J. Information Resources Management Journal. 10(1), 6–25 (1997)
3. Bhanu, R., Raghunathan, T.S.: Research Report—Adaptation of a Planning System Success Model to Information Systems Planning. J. Information Systems Research. 5(3), 326–340 (1994)
4. Jerry, C.C., William, R.K.: Measuring the Performance of Information Systems: A Functional Scorecard. J. Management Information Systems. 22(1), 85–115 (2005)
5. Xiao, J., Cui, Y., Wang, J., Luo, F., Li, Y., Wang, S., Wang, H.: A Hierarchical Performance Assessment Method on the Distribution Network Planning. J. Automation of Electric Power Systems 32(15), 36–40 (2008) (in Chinese)
6. Ye, Y.S., Wang, J.Z.: Research on Evaluation Indicators System for Whole Life Cycle of Information System. J. Computer and Modernization (199), 62–65 (2012) (in Chinese)
7. PDCA - Wikipedia, the free encyclopedia, http://en.wikipedia.org/wiki/PDCA

A New Look into Web Page Ranking Systems

Thi Thi Zin[1], Pyke Tin[2], Hiromitsu Hama[3], and Takashi Toriu[2]

[1] Faculty of Engineering, University of Miyazaki, Miyazaki, Japan
[2] Graduate School of Engineering, Osaka City University, Osaka, Japan
[3] R&D Center of 3G Search Engine, Incubator, Osaka City University, Osaka, Japan
thithi@cc.miyazaki-u.ac.jp

Abstract. This paper proposes a new way of looking into Web page ranking systems by using some concepts of queuing theory in operations research and stochastic water storage theory in hydrology. Since both theories queuing and stochastic water storage are rich in technology as well as application aspects, the new look in this paper may lead to new directions in Web page ranking systems and related research areas. In doing so, first this paper draws some analogies between a Web page ranking system and theory of queues. Then it shows how a Web page ranking system can be tackled to reduce current obstacles by using queuing theory techniques. In the second, a Web page ranking system is modeled as a framework of stochastic water storage theory to derive a list of Web page rankings. Third and finally, the outcome results of rankings obtained by using the proposed two theories queuing theory and stochastic water storage are compared and analyzed analytically as well as experimentally. The experimental results show the proposed new look is promising for establishing a new research area which can improve the current situations and difficulties occurred in search engines and their ranking systems in particular and some problems in World Wide Web as a whole.

1 Introduction

It has been a little over one and half decades after the two most popular web page ranking systems Google's Page Rank [1] and Kleinberg's Hyperlink-Induced Topic Search (HITS; also known as hubs and authorities) [2] appeared in science literature. Both algorithms are based on hyperlink analysis on Web page graphs. In particular, the key idea is to consider hyperlinks from a source page to sink pages as an influential factor on ranking sink pages. Based on this concept a lot of excellent research woks by various researchers have been proposed to improve the existing web page ranking systems [3-6]. But different methods have utilized different concepts to analyze Web link graphs [7-8]. For example Google`s Page Rank model assumes that the hyperlinks between the Web pages can be approximated as the transition probabilities of a Markov Chain of lag 1[9]. On the other hand HITS algorithm observes a Web graph as bipartite forms [2, 10]. Even more some researchers have introduced community ranking systems by using social network platforms based on popularity, reliability and relevancy measures [11-12]. Thus, it is learnt that various approaches have

H. Sun et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 329, DOI: 10.1007/978-3-319-12286-1_35

been appeared to measure the Web Graph by using different aspects of measures such as relevancy, reliability and popularity concepts. The success of Google has made the role of Page Rank significance as a metric measure for Web pages. This has also led to a variety of modifications and improvisations of the basic PageRank metric. These have either focused on changing the underlying model or on reducing the computation cost.

On the other hand the rise of giant social networks such as Facebook, Twitter and YouTube and so on has influenced on the ways how users are seeking information. In this aspect, some researchers have introduced a concept of social community ranks and combined them with contents of web pages to form an integrated ranking system [13-14]. But they have said that even though their approach is new, it is only at infant stage so that it remains more research has to be done in this direction. However, the original Page Rank algorithm is quite popular and has led many excellent research works, the authors of this paper feel that it is worthwhile to look into to the Page Rank algorithm from new perspectives. Therefore this paper explores a new look into the original Page Rank algorithm from the perspectives of queuing theory and Dam theory which had been introduced by P.A P Moran [15] later extended by J. Gani [16] and R. M. Phatarfod [17].

2 Briefly about Web Page Ranking System

As a theoretical foundation it is worthwhile to realize that an overall ranking of a web page for a given query can be incorporated by a probability distribution [18-19]. To begin with, the probability of a document d is relevant for a given query q, $\Pr.(d|q)$ is expressed by using Bayes' theorem as

$$Pr.(d|q) = Pr.(d).Pr.(q|d) \,/\, Pr.(q). \quad (1)$$

From web page ranking perspective views, since the term $Pr.(q)$ does not depend on the document d, the term $Pr(q|d)$ becomes one of the central roles for the information search engines to evaluate the importance of a web page in the web graph. Also this simple formula had led to the most popular web page ranking system, Page Rank algorithm and variety of other ranking systems for information search systems.

As a starting point, the simplest version of Page Rank, so called standard Page Rank which had made Google success is defined as the stationary distribution of a random walk in the web graph. Specifically, $PR(i)$ denotes the value of web page i for i= 1,2,…,N with N as the total number of pages in the web graph. $Pr.(j|i)= p_{ij}$ is the probability that the walker jumps from page i to page j and W_j is the set of pages that points to page j for j=1,2,…,N and W denotes the set all pages. Then the fundamental equation for $PR(i)$ can be described as shown in (2).

$$\mathrm{PR(j)} = (1-\alpha) \sum_{i\in Wj} p_{\mathrm{ij}}\mathrm{PR(i)} \;+ \alpha \sum_{i\in W} PR(i) \,/N. \quad (2)$$

Traditionally the value of α is chosen as 0.15, and it appears that this value provides reasonable ranking for the web pages.

In order to write the equation (2) in matrix form, first organize a matrix H=[h_{ij}] where h_{ij} =(1- α) p_{ij} + α/N for $i \in W_j$ and h_{ij} = α/N for $i \notin W_j$.. Again, an identity matrix I is defined as a diagonal matrix whose diagonal elements are all ones and the remaining elements are all zero.

Then the equation (2) can be written as

$$\pi = \mathrm{G}\,\pi\,, \qquad .(3)$$

where π = [PR(1), PR(2),.....PR(N)] $^{\mathrm{T}}$ is a column vector and G= [H+ α I].

The matrix G= [H+ α I] is known as Google matrix. This matrix is stochastic (each row sums to 1), irreducible (all pages are connected due to the teleportation jump), a periodic (G_{ii} >0), and primitive (G^k > 0), which implies that a unique positive π exists and power method guarantees to converge to this vector. Given some initial distribution π (0), e.g; π (0) = e, the power method is defined as an iteration procedure:π(k) = G π(k–1) , $k \geq 1$.

Note that the limiting distribution does not depend on the initial distribution π(0). In the framework of stochastic processes one can view the Page Rank of a random web page at time t as random variable R_t. Then $PR(j)$ for j= 1,2,….,3 can be interpreted as the probability of R_t =j. Then the equation (2) becomes a stochastic equation which has been occurred in queues, dams and branching process theories.

One of the objectives of this paper is that it provides better understanding on what the terms $Pr.(d)$, $Pr.(q|d)$ might in equation(1) and the terms $PR(j)$ for j=1,2,…,N look like, and how they related to theory of queues and theory of water storage for opening up a new research in web search engine frameworks. To achieve this goal, the importance of a web page will be defined for a search engine based on convex combination of rankings according to different criteria, such as relevancy, reliability and popularity of the pages in the web graph.

In particular the remaining part of this paper is organized as follows. In section 3, the analogies between Web Page Rank formulation, queueing theory concepts and theory of water storage systems are presented. Then a new look at Web Page Rank System is introduced and the quantities of interests are derived in section 4. Some simulation works are shown in section 5 followed by concluding remarks in section 6.

3 Analogies between Web Page Rank, Queueing Theory and Water Storage Systems

In the analysis of many queueing systems or water storage systems (dams) the following Random Walk type of recurrence relationship is fundamental and plays an important role.

$$W_{t+1} = \max\,[W_t\ +1\text{-}X_t\,,\,0],\ \text{for}\ t= 1,2,3,\ldots\ . \qquad \ldots(4)$$

There are a variety of ways to interpret and make assumptions for the processes {W_t} and {X_t }. For example in theory of queues, X_t represents the arrival patterns process for new customers coming into a queue along with other processes such as service time distribution and service mechanism. On the other hand, from the dam theory aspect, the random variable X_t is the amount of water inflows into a dam during a time interval (t, t+1) for t=0,1,2,……. . Keeping these concepts in mind, the random variable X_t shall be considered as the amount of incremental ranks added to a random web page due to in-degrees. In all cases, W_t will be the number of customers who are waiting for service including the one being served. while it will represent the depletion from the maximum level, say K-1 where K is the capacity of a finite dam. In terms of web rank, K can be considered as the total number web pages N in the graph and W_t will represent the level of page ranks at time t. It is this analogy to be exploited in analyzing Web Page Ranking Systems. Since the queueing and dam theory have a variety of analytical results for W_t by taking independent and dependent structure of input process into account, a new way of formulating Web Page Ranking will be explored and examined in this paper. The overview of proposed system is described in Fig 1.

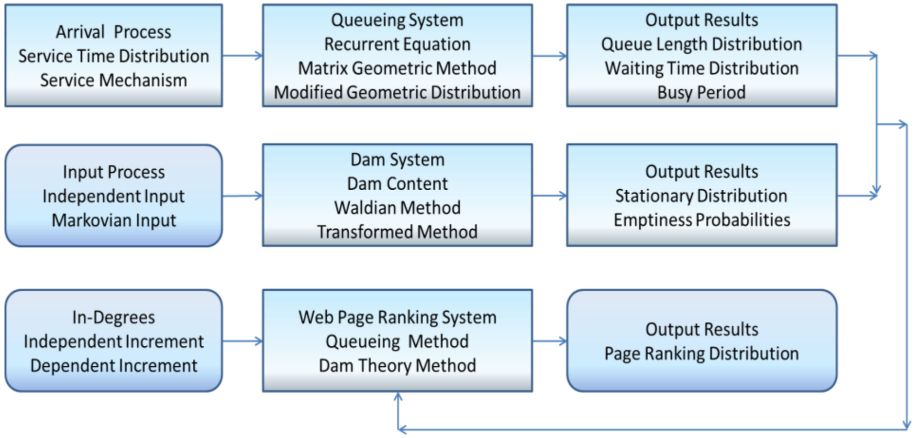


Fig. 1. Overview of Proposed System

3.1 Analysis of Web Page Ranking System in the Frameworks of Queues and Dams

In this subsection a new Web Page Ranking System is analyzed by using the techniques developed in queue and dam theories. For this concern, two cases will be considered by introducing concepts of both independent type and Markov type in-decrees.

A. Independent Type In-Degree Case

In this case, { X_t } forms a sequence of independent and identically distributed random variables in queues and dams while it will represent rank increments due to in-degrees pointing the page considered. It is also assumed that the probability generating function the random variable X_t is $g(s) = \Sigma \text{ Prob}(X_t=j)\ s^j$ sum over all j.

Generally speaking, a problem of interest to hydrologists, queue controllers and web search engines optimizers is finding the limiting distribution of the process $\{W_t\}$.

Let $Pr\ (W_t = j) = \pi_j$ when t tends to ∞. In queues and dams, when $\mu = g'(0) > 0$ the limiting distribution π_j is given by

$$\pi_j = (1\text{-}\theta)\ \theta^j, \tag{5}$$

where θ is the smallest positive root of $g(s) = s$.

This leads to the results

$$\pi_j \cong (1\text{-}\theta)\ \theta^{K\text{-}1\text{-}j} / 1\text{-}\theta^K \quad \text{for } j=0,\ldots.,K\text{-}1. \tag{6}$$

The probability in (6) is generally used to estimate the size of a dam for reliability factor P as shown in (7).

$$K \cong 1 + (\log P) / \log \theta. \tag{7}$$

By using the analogy between Dams and Page Rank, the Page Rank of a random page can be approximated as

$$PR(j) \cong (1\text{-}\theta)\ \theta^{N\text{-}1\text{-}j} / 1\text{-}\theta^N, \text{ for } j=0,\ldots..,N\text{-}1, \tag{8}$$

where the number of pages in the web graph is estimated for reliability factor P as

$$N \cong (1\text{-}\theta)\ \theta^{N\text{-}1\text{-}j} / 1\text{-}\theta^N\ . \tag{9}$$

B. Markov Type In-Degree Case

In queueing theory framework, it has been shown that the queue length process embedded in a general arrival process such as moving average [20] and Gamma Markov [21] has a modified geometric distribution similar to those described in (6) and (7).

Let the sequence $\{X_t\}$ a Markov chain with transition probability q_{ij} for i, j 0,1,2,….,N-1. Let $Z_{ij} = s^i q_{ij}$ and $[Z^n_{ij}] = [Z_{ij}]^n$. Then it can be proved that $\lambda(s) = \lim\ (Z^n_{jj})^{1/n}$ when n tends ∞ exists and independent of j. By assuming that $0 < \theta < 1$ is a solution of $\lambda(s) = s$, the limiting distribution of $\{W_t\}$.

Moreover for the case of special type of Markov Chain such that

$$\sum q_{ij}\ s^j = B(s)\ [A(s)]^i\ , \tag{10}$$

where $B(s)+[1+a(1\text{-}\rho)–a(1\text{-}\rho)\ s]^{-1}$ and $A(s) = \frac{[\,(1+a)(1-\rho)-(a\ (1-\rho)-\rho)s]}{1+a(\ 1-\rho)-a(1-\rho)s}$

$0<\rho<1$, is the correlation coefficient of lag 1 and a is the mean value of $\{X_t\}$. Pakes And Phatarfod [22] had proved that the limiting distribution of $\{W_t\}$ is as shown in equation (11).

$$\pi_0 = \lim_{t \to \infty} \Pr\ (Wt = 0) = 1 - \frac{1+q}{1+a}\ , \tag{11}$$

$$\pi_j = \lim_{t \to \infty} \Pr\ (Wt = j) = \frac{1-q2}{1+a}\ \ \mathrm{q}^{j\text{-}1} \quad \text{for } j \geq 1. \tag{12}$$

With $q = (2\rho)^{-1}[-a(1\text{-}\rho) + (\ \{a(\ 1\text{-}\ \rho^2\ )\ \}^2\ +4\rho)^{1/2}\ )\ ]$. By using equations (11) and (12), the page rank distribution is given by

$PR(0) = [a\text{-}q]/[(1+a)\text{-}(1+q)\ q^{N\text{-}1}],$
$PR(j) = (1\text{-}q^2\ )\ q^{\ N\text{-}j\text{-}1} / [(1+a)\text{-}(1+q)\ q^{N\text{-}1}]$ for j= 1,2,….,N-1.

4 Some Simulated Results

In this section some simulated results for the distribution of web page rank is analyzed. The relationship between the correlation of the web pages and the number pages within a dataset is also discussed by using numerical methods. Although most of existing page rank algorithms are subjects of iterations demanding the rate of convergence, the proposed methods needs only one time computation. So it is not necessary to worry about the convergence problems. The proposed method has taken care of convergence concepts implicitly. This merits to reducing the computation time for calculating Web Page Ranks. The rank distribution induced using queuing and dam theory concept also can measure the correlation between PageRank and the In-Degree. In this aspect, it is shown that how much one can vary the values of the number of in-degree and the maximum correlation one can use is presented. In Table 1, for small value of number of pages, the results of page rank distribution is presented for large correlation value and small size of in-degree. In Table 2, the results of rank distribution for minimum allowable correlation and the maximum size of in-degree are shown. In similar way, for large value of number pages within the dataset, the effects of correlation and in-degree are presented in Table. These results are also shown in graph forms in Fig 2. Again, through the varying the parameters correlation, in-degrees and the number of pages it is observed that the correlation coefficient value is range from 0.7 to 0.98 for large value of *N*, the total number of pages within the datasets. For small *N*, the correlation cannot exceed 0.85 but in-degree is the same as in the case of large value of *N*.

Table 1. Distribution of Ranks for Small *N*=10

No	Pages	Rank for ρ=0.85, a=7	Rank for ρ=0.85, a=6	Rank for ρ=0.85, a =5	Rank for ρ=0.85, a =4	Rank for ρ=0.85, a =3	Rank for ρ=0.85, a =2	Rank for ρ=0.75, a =7	Rank for ρ=0.65, a =7
1	π0	0.905665	0.848207	0.796386	0.742748	0.681991	0.616825	0.634724	0.823814
2	π1	0.008768	0.012517	0.015546	0.01933	0.025397	0.035363	0.034334	0.018587
3	π2	0.027446	0.027446	0.027446	0.027446	0.027446	0.027446	0.027446	0.027446
4	π3	0.009557	0.014411	0.018517	0.023186	0.029659	0.038673	0.037228	0.01907
5	π4	0.009978	0.015462	0.02021	0.025393	0.032051	0.040442	0.038765	0.019316
6	π5	0.010417	0.01659	0.022057	0.027811	0.034636	0.042292	0.040366	0.019565
7	π6	0.010876	0.017801	0.024073	0.030458	0.037429	0.044227	0.042032	0.019818
8	π7	0.011355	0.0191	0.026273	0.033358	0.040448	0.046251	0.043768	0.020074
9	π8	0.011855	0.020493	0.028675	0.036533	0.04371	0.048367	0.045575	0.020333

Table 2. Rank Distribution for Large *N*

No	Pages	Rank for ρ=0.98, *a* =8	Rank for ρ=0.98, *a* =7	Rank for ρ=0.98, *a* =6	Rank for ρ=0.98, *a* =5	Rank for ρ=0.98, *a* =4	Rank for ρ=0.98, *a* =3	Rank for ρ=0.98, *a* =2	Rank for ρ=0.7, *a* =2
1	π0	0.784292	0.756428	0.720546	0.672638	0.605487	0.504661	0.336483	0.34293
2	π1	0.012647	0.012525	0.012247	0.011729	0.010824	0.009235	0.006269	0.018917
3	π2	0.011906	0.011881	0.011710	0.011309	0.010527	0.009062	0.006210	0.018372
4	π3	0.011208	0.011270	0.011197	0.010904	0.010238	0.008893	0.006151	0.017844
5	π4	0.010550	0.010691	0.010706	0.010513	0.009957	0.008728	0.006093	0.017330
6	π5	0.009932	0.010141	0.010237	0.010136	0.009684	0.008565	0.006035	0.016831
7	π6	0.009350	0.009619	0.009788	0.009773	0.009418	0.008405	0.005978	0.016346
8	π7	0.008801	0.009125	0.009359	0.009423	0.00916	0.008249	0.005922	0.015876
9	π8	0.008285	0.008656	0.008949	0.009085	0.008909	0.008095	0.005866	0.015419
10	π9	0.007800	0.008210	0.008557	0.008760	0.008664	0.007944	0.005810	0.014975

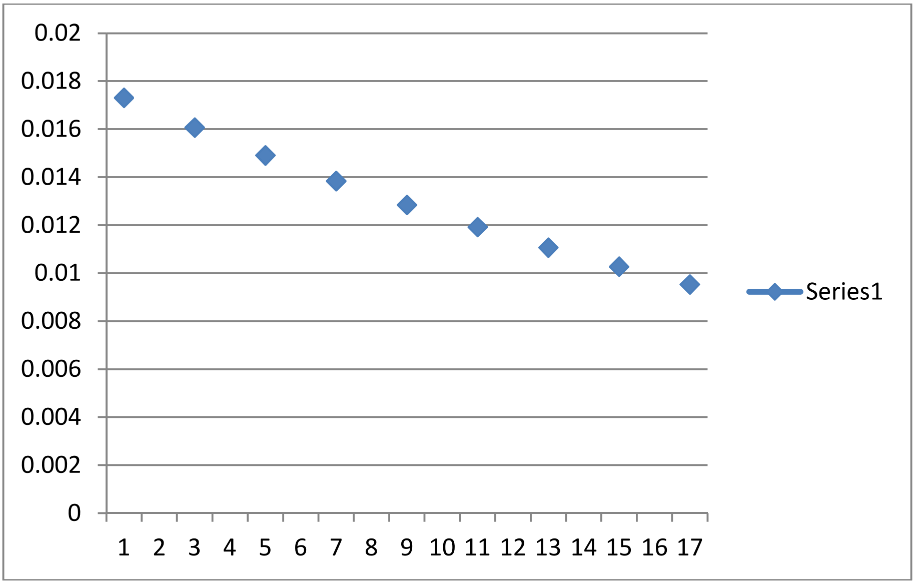


Fig. 2. For small *N*=10, ρ=0.85 and In-degree=7

5 Conclusions

A new look into Google's Page Rank algorithm has been proposed in this paper. The use of queue concepts and dam theory results makes the proposed method promising. Although an empirical evaluation was done using synthetically generated data, the results show that the proposed method provides a variety of result rankings. In future, focus will be made on performing evaluations using real world data.

References

[1] Brin, S., Page, L.: The anatomy of a large-scale Hyper-textual Web Search Engine. Computer Networks 30(1-7), 107–117 (1998)

[2] Kleinberg, M.: Authoritative Sources in Hyperlinked Environment. In: 9th Annual ACM-SIAM Symposium onDiscrete Algorithms, pp. 667–668 (1998)

[3] Desikan, P., Pathak, N., Srivastava, J.: Incremental Page Rank Computation on Evolving Graphs (2005), `http://www-users.cs.umn.edu`

[4] O'Madadhain, J., Smyth, P.: EventRank: A framework for ranking time-varyingnetworks. In: LinkKDD, pp. 9–16. ACM (2005) (Cited on pages 10 and 16)

[5] David, F., Ryan, G., Rossi, A.: A Dynamical System for PageRank with Time-Dependent Teleportation. Purdue University (November 20, 2012)

[6] Eiron, N., McCurley, K., Tomlin, J.: Ranking the Web Frontier. In: Proc. 13th Conference on World Wide Web, pp. 309–318. ACM Press (2004)

[7] Richardson, M., Domingos, P.: The intelligent surfer: Probabilistic combination of link and content information in pagerank. In: Advances in Neural Information Processing Systems (2002)

[8] Haveliwala, T.: Topic-Sensitive PageRank. In: Proceedings of the Eleventh International World Wide Web Conference (May 2002)

[9] Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank Citation Ranking: Bringing Order to the Web. Bringing Order to the Web. Technical Report. Stanford Digital Library Technologies (January 1996), `http://www-diglib.stanford.edu/diglib/pub/`

[10] Ding, C., Zha, H., He, X., Husbands, P., Simon, H.D.: Link Analysis: Hubs and Authorities on the World Wide Web (May 2001) LBNL Tech Report 47847 (1998)

[11] Tin, P., Zin, T.T., Toriu, T., Hama, H.: A Cluster Based Ranking Framework for Multi-Typed Information Networks. In: Accepted in IIh-MSP 2014, Kita Kyusyu, August 27-29 (2014)

[12] Zin, T.T., Tin, P., Toriu, T., Hama, H.: A Stochastic Model for Measuring Popularity and Reliability in Social Network Systems. In: SMC 2013, pp. 462–467 (2013)

[13] Tin, P., Zin, T.T., Toriu, T., Hama, H.: An Integrated Framework for Disaster Event Analysis in Big Data Environments. In: IIH-MSP 2013, pp. 255–258 (2013)

[14] Hama, H., Zin, T.T., Tin, P.: A Hybrid Ranking of Link and Popularity for Novel Search Engine. International Journal of Innovative Computing, Information and Control: Special Issue on Multimedia Processing and Network Technologies 5(11(B)), 4041–4049 (2009)

[15] Moran, P.A.P.: A probability theory of dams with a continuous release. Quart. J. of Math. 7, 130–137 (1956)

[16] Gani, J.: Problems in the probability theory of storage systems. J. Roy. Statist. Soc. B 14, 181–207 (1957)

[17] Tin, P., Phatarfod, R.M.: On Infinite Dams with Inputs Forming A Stationary Process. J. Appl. Prob. 11, 553–561 (1974)

[18] Pennock, D.M., Flake, G.W., Lawrence, S., Glover, E.J., Giles, C.L.: Winners don't take all: Characterizing the competition for links on the Web. Proceedings of the National Academy of Sciences 99(8), 5207 (2002)

[19] Kraaij, W., Westerveld, T., Hiemstra, D.: The importance of prior probabilities for entry page search. In: SIGIR 2002: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 27–34. ACM, New York (2002)

[20] Finch, P.D., Pearce, C.: A second look art a queueing system with moving average input process. J. Aust. Math. Sot. 5, 100–106 (1965)

[21] Tin, P.: Some Problems on Dams and Queues with Correlated Inputs. Dissertation, Monash University (1976)

[22] Pakes, A.G., Phatarfod, R.M.: The Limiting Distribution for Infinitely Deep Dam with A Makovian Input. Stochastic Processes and Their Applications, 199–209 (1978)

Seismic Qualification of Telecommunication System in Nuclear Power Plant

Zhichun Li and Yong Zheng

CGN Power Co., Ltd., Centre of Information Technology, ShenZhen, China
{lizhichun,zhengyong}@cgnpc.com.cn

Abstract. To guarantee the effective communications under earthquake and enhance the utility of communication system in nuclear power plants, it is proposed to do seismic qualification test for all telecommunication system in nuclear power plant. In this paper, a principle include object, scope method and setup of the qualification is provided. It is found that if telecommunication equipments pass such qualification test, the availability of communication system and the safe operation of nuclear power plant can be guaranteed when earthquake occur.

Keywords: Nuclear power plant, seismic qualification, earthquake, telecommunication system.

1 Introduction

Safety is the most important prerequisite for nuclear power plant. After the Fukushima Nuclear Power Plant disaster, the effects of earthquake to nuclear safety get more and more attention. And in the design of nuclear power plant, the earthquake has to be taken into account.

The telecommunication system is not nuclear safety relevant system, but telecommunication equipments distributes everywhere in nuclear power plant, so they maybe nearby a nuclear safety relevant equipment and may affect the nuclear safety relevant equipment. So there is also seismic requirement for the telecommunication system. There is not function request during and after the earthquake, but there is integrity request. We must ensure the structural integrity of the telecommunication equipments during and after exposure to the dynamic effect of seismic load, there shall be no collapse, missile, falling off of the structure and so on.

To verify the structural integrity during certain seismic load, there are two methods: calculating method or testing method. Because the telecommunication equipments have very complex structure and material, there will be bigger deviation in modeling and the calculating. In this paper, a testing method is discussed and demonstrated to be available in verifying the structural integrity during certain seismic load.

In this paper, we deeply analyzed the principle to choose test specimens, the test method and the acceptance criteria and all the principles to do the seismic qualification.

H. Sun et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 329, DOI: 10.1007/978-3-319-12286-1_36

Through the seismic qualification, the structural integrity under certain earthquake load can be guaranteed, and avoid the destroying to the nuclear safety.

2 Test Specimen

Telecommunication system in nuclear plant includes the Public Address System, Siren Alarm System, Intercom System, Telephone System, Direct Phone System, Secure Telephone System, Clock System, premises distribution system and CCTV System. All of them must be qualified.
All the equipments can be classified into two kinds:

(1) Central unit in communication room and their supporting equipment;
(2) Connection boxes and terminals of all sub-systems

For the two types of equipments, there shall be different specimen choosing principle.

2.1 Central Unit

For the communication central cabinet, there will be hundreds of combination for different usage of central unit, it's impossible to test every combination of theses equipments and cabinet. So we have to choose a most severe configuration cabinet to do the test, and the test result can be appropriated for all cabinet. How to choose a most severe configuration cabinet? Here are the principles:

In the test, all substitutes of components will be installed from the top of the cabinet except the UPS and batteries, Batteries and UPS will always installed in the bottom of the cabinet. Batteries and UPS have the biggest density.

All substitutes of components will be installed in the cabinet following these specifications:

(1) Batteries and UPS will always be installed in the bottom of the cabinet.
(2) Equipment will be installed from the top to the bottom of the cabinet.
(3) The components with bigger density will be installed upon the components with smaller density.

All real cabinet on site is designed following these specifications:

(1) Batteries and UPS will always be installed in the bottom of the cabinet
(2) Equipment will be installed from the bottom to the top of the cabinet.
(3) The components wit bigger density will be installed under components with small density.

If follow these specifications, the gravity center of the real cabinet on site will be lower than the tested one, so the tested cabinet is the most severe configuration. The specimen cabinet is the worst case covering all kind of cabinets actually installed on site. The test result is appropriate for cabinet which is equipment installed on site.

2.2 Junction Box

The junction box is composed of box and the equipments inside of the box. There are several types of box, some are made of polycarbonate and some are made of cold rolled steel. The equipments inside of the box are airlock, terminal block or wiring module. Due to HAF J0053, we should select the harshest equipment to test, so for each size of box, we select the one who have heaviest equipments in to test. And the result of test is appropriate for connection boxes which use the same box. Base frame of the connection boxes is not used during the test.

2.3 Terminal Equipment

For the terminal equipments (loudspeaker, siren alarm, intercom station, operation console, telephone, camera, computer and wireless access point), typical specimens shall be chosen for the test, here is the principle of choosing specimens:

(1) For different terminal equipment, all shall be tested to verify their stability during the seismic

(2) For similar equipments which are the same brand, if they have similar structure and similar installing method. The heavier equipments should be chosen for the test, and the result of test is also appropriate for the lighter ones.

(3) The installing method of the test specimens during the test should the same as the installing method on site except the anchor bolts. If an anchor bolt is used on site, a same size and same rigidity ordinary bolt can simulate the anchor bolt during the test.

3 Test Method

The specimens will be mounted on the table in a manner that simulates the intended service mounting. The mounting method of the specimen to the support or mounting steel plate will be the same as that recommended for actual service and will use the recommended bolt size, type torque.

Multi-frequency method is used in the seismic qualification test. Artificial time history shall be inputted along three orthogonal axes of the specimen to fulfill five OBE and one SSE seismic tests. OBE and SSE seismic simulation tests are carried out by controlling the acceleration of the seismic table as the control signal.

3.1 Condition Inspection before Test

Before the test, all status of the specimens should be checked and recorded. There shall not be any loose. All components shall be installed exactly.

3.2 Dynamic Characteristics Detect Test

According to section 6.2.8 of IEC 60980-1989, apply the sinusoidal excitation with frequency sweep along each orthogonal axis of the specimens respectively. The sinusoidal excitation amplitude is 0.2g, frequency range is 1Hz to 55Hz and the sweep speed is 1 octave/minute.

PARAMETER TABLE

Excitation direction	X/Y/Z
Excitation type	Sine sweep
Sweep rate	1oct/min
Acceleration	0.2g
Frequency range	1Hz to 55Hz
Duration time	120s

Fourier transfer function is used to calculate resonance frequency, which the time history of the control point is the input parameter and the time history of top of the specimens are the response parameters.

3.3 Load of Seismic Qualification Test

Artificial time history is generated based on the RRS (Required Response Spectrum) of corresponding Nuclear Power Plant. Duration of the time history is 30 seconds. The tests using three-axis method are controlled with a digital control using the simulated time-histories. The simulated time-histories must produce TRS (Target Response Spectrum) to envelop the RRS. The time history is used to excite the specimens during the test.

The analysis of the TRS and PSD of the control point time history after the test is to validate program realization. The TRS must equal or more than the RRS at all points above 3.5Hz. The TPSD must equal or more than the 80% RPSD at all points above 3.5Hz.

*the start frequency 3.5 Hz is accord to the section 6.2.9 of [4].
*the RPSD and TPSD are computed based on RRS and time history.

3.4 OBE Seismic Qualification Test

According to section 6.2.9 of IEC 60980-1989, five time history OBE (Operational Basis Earthquake) level tests shall be carried out along three orthogonal axes of the specimen. The duration of each OBE test is 30 seconds.

The acceleration signal shall be recorded during the test. The structural integrity of the specimen shall be inspected during and after each test.

3.5 SSE Seismic Qualification Test

After five OBE seismic simulation tests, one SSE (Safe Shutdown Earthquake) test shall be applied along three orthogonal axes of the specimen. The duration of SSE test is 30 seconds too.

Between OBE and SSE test, the minimal interval between each operation is 3 minute.

OBE test is half of SSE.

The acceleration signal shall be recorded and the structural integrity of the specimen shall be inspected during the test.

3.6 Condition Inspection after Test

The structural integrity of the specimens shall be inspected again, in order to ensure the specimen be in good condition after test.

4 Acceptance Criteria

The telecommunication system is not a nuclear safety related system, so there is not function request during and after the earthquake, but there is integrity request. So we have to check the integrity of tested specimens, the acceptance criteria is as follow:

(1) There shall be no collapse of the tanks, and the bolts shall be no disconnection.

(2) There shall be no falling off of other specimens, and the bolts shall be no drop.

5 Conclusion

A method of seismic qualification for telecommunication system in EPR nuclear power plant is studied in this paper. The method specify the principle of test specimens choosing; provide the method of load calculation during seismic qualification test, specify all the test steps clearly, and specify the acceptance criteria. After all, the paper specifies every aspect of the seismic qualification.

Through the seismic test, the integrity of telecommunication system during and after the earthquake can be proved, and the safety of nuclear power plant can be guaranteed.

Acknowledgment. The authors would like to thank anonymous reviewers for their constructive comments, which improved quality of the paper.

References

1. Hargreaves, S.: Half of U.S. nuclear reactors over 30 years old (N/OL). CNN Money (March 15, 2011), `http://money.cnn.com/2011/03/15/news/economy/nuclear_plants_us/index.htm`
2. Andrews, A., Folger, P.: Nuclear power plant design and seismic safety. Considerations, Congressional Research Service Report for Congress. R41805 (2012)
3. Cornell, C.A.: Engineering seismic risk analysis. Bulletin of Seismic Society of America 58, 1583–1606 (1968)
4. KTA 2201.4 Design of Nuclear Power Plants against Seismic Events
5. CRT91C 11200 Seismic resistance of equipment Generic provisions for time history-based biaxial test
6. McGuire, R.K.: Fortran computer program for seismic risk analysis. U. S. Geological Survey, Open File Report, 67–76 (1976)
7. STUK – Radiation and Nuclear Safety Authority. Maanjaristysten huomioon ottaminen ydinvoimalaitoksissa. OHJE YVL 2.6/19 (2001)
8. Gasparini, D., Vanmarkcke, E.: Simulated earthquake motions compatible with prescribed response spectra. MIT Department of Civil Engineering Research Report (1976)
9. Safety Guide on Earthquakes and Associated Topics in Relation to NPP Siting, HAF0101. Approved jointly by NNSB and SSB. A collection of safety guides for NPP, NNSB (1992)
10. ASCE STANDARD 4-86. Seismic analysis of safety-related nuclear structures and commentary on standard for seismic analysis of safety related nuclear structures (1996)
11. U.S. Atomic Energy Commission, Regulatory Guide 1.60. Design response spectra for seismic design of nuclear power plants. Version 1 (1973)

Part VI
Technologies for Next-Generation Network Environments

A Virtual Network Guard System Based on Cloud Computing Environments

Bing-Zhe He[1], Kuan-Ling Huang[1], Hung-Min Sun[1], and Raylin Tso[2]

[1] Department of Computer Science, National Tsing Hua University, Hsinchu,
Taiwan, R.O.C.
{ckshjerho,kent}@is.cs.nthu.edu.tw, hmsun@cs.nthu.edu.tw

[2] Department of Computer Science, National Chengchi University,
Taipei, Taiwan, R.O.C.
raylin@cs.nccu.edu.tw

Abstract. The cloud computing is one of the most popular issues in recent years. Many service providers have provided the cloud solution using virtualization such as Amazon EC2. We are facing the new threats in the virtual environment. Since the virtual network is different to the traditional environment, we have to face with new threats that do not exist in the traditional network environment. In this paper, we provide a solution Virtual Network Guard System (VNGS) to solve the problems in virtual network that we face in the virtual environment. We modify the network interface controller in the virtual environment to limit the guest operation system access rights. We also provide a centralize management server to deploy filtering processes and collect the alert information. Finally, we evaluate the performance of our system with normal network interface controllers, and the results shows that the performance is acceptable.

Keywords: Cloud Security, Packet Filter, Virtual Network Security System.

1 Introduction

Nowadays, the virtualization technology is used to reduce the number of physical machines in enterprises and make our life easier and more efficient. However, the security issue is an important challenge when an IT manager considers to distribute virtualization. We are facing new threats in virtual environments and some attacks have targeted on the virtual environment [7].

Since the virtual network environment is different to physical network environments, a virtual environment can make some network attack strategies be triggered easily or have some new threats. However, seldom researches have focused on the security of the virtual machines. Some of them are limited by the special hardware or architecture [9,5]. Few of them discuss the network security of virtual environments.

In many existing network attack strategies, using a raw socket to create fake network packets is the core technology for an attacker to launch a network attack,

H. Sun et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 329, DOI: 10.1007/978-3-319-12286-1_37

such as ARP spoofing. The ARP spoofing attack sends packets with fake MAC addresses. This can make all the machines which are located in the same local area network (LAN) believe that the attacker is another machine. In normal network environments, it is hard for an attack to infiltrate into the same LAN with targets. Nevertheless, virtualization can make the attacker and targets both be in the same LAN, and the attacker can launch the malicious attacks to other targets.

In this paper, we propose a solution called Virtual Network Guard System (VNGS) to solve the security problems in virtual network environments. Our goal is to provide a solution that an administrator can easily limit the network access rights for each VMs in the cloud environment. By modifying virtual network interface controller to filter illegal network packets such as fake MAC addresses and IP addresses, the cloud administrators can implement an easy filtering process to control virtual network interface. In this way, our solution can reduce complexity of managing virtual machines in the cloud. Finally, we also evaluate the performance of our architecture.

2 Background

In this section, we review some researches which are pertinent to the problem of the network security in cloud computing environments.

2.1 Enhance Security of VMM

Terra [3] is a virtual machine-based platform, and provides a new architecture of virtual machine based on the special hardware device. This system uses a tamper-resistant hardware platform to equip with trusted virtual machine monitor (TVMM). It provides a simple and flexible programming model that allows application designers to build secure applications in the same way they would on a dedicated closed platform. However, this solution is limited by special hardware devices and the VMM has to be redesigned. The paper [4] provides a VM introspection based architecture for instruction detection called VMI. VMI is a virtual machine monitor embedded with IDS. It not only focus on protecting network, but also provides the guest OS. The VMI is modified by VMware Workstation for Linux.

2.2 Enhance Network Environment

To protect the network in the cloud computing environments, it is a good way to split the network as many virtual local area network(VLAN). In physical network environments, it requires some advanced network device to support this function. Open vSwitch [8] is a network switch sepcifically built for virtual environments. By spliting the VLAN, we can build several isolated LANs. It can prevent attackers sniff the virtual network to steal data from other gesut OS. However, the performance of open vSwitch still can be improved.

Table 1. The comparison between those solutions and our system

Name	Advantage	Disadvantage
TVMM	hardware protect and high performance	require special hardware support
VMI	enhance VMM security and load customized plug-in	low performance
OpenvSwitch	isolate virtual network environments	low performance
VNGS	load customized filter and high performance	packets filtering only

2.3 Comparison

To compare the advantage between those solutions and our system. Table 1 shows the comparison. TVMM provides high performance and hardware guarantee, but it requires a special hardware device. For an administrator in a data center, it is hard to embed TVMM hardware to all the devices in the data center. VMI provides a redesigned VMM architecture embeds IDS functionality. VMI can monitor the behavior of guest OS and load customized detection plug-in. However, the performance is not good and it requires redesign VMM. OpenvSwitch is a virtual network switch which emulates the behavior of physical network switch. It can isolate the virtual network just like a physical network switch does. But the software emulating causes low performance. In our VNGS, it provides a administrator to load a customized filtering process to filter network packets. We only modify the source code of virtual NIC and the performance analysis in 5 shows that VNGS does not increase many overhead.

3 Attack Model

Now, we describe an attack scenario based on virtual network environments with QEMU virtual machine. In physical network environments, the conditions of the ARP spoofing attacks are not easy to reach since the attacker and target host have to be connected under the same local network. For some special targets, it is hard for the attacker to infiltrate into the same LAN. For example, a target host is under a private local area network or attacker and target host are in different countries. If they are not in the same LAN, the ARP packets can not be diverted to target host. If the ARP packets can not be diverted to target host, the attack can not work successfully.

Nevertheless, if there are no any advanced configurations, all guest OSes are allocated in the same local network as default. In a data center, there can be a lot of host machines which operate a virtual machine monitor. Each host can operate more than one guest OSes. If an attacker can occupy one of these guest OSes, the attacker may obtain enough information to launch ARP spoofing attack. Not only ARP spoofing attack but other network attacks are achieved easily when attackers and victims are in the same local network.

To model the attack in virtual network environment, we make some assumptions as follow:

- Attackers have the ability to occupy the guest OSes and send modified instructions to the host.
- Virtual machine monitor does not have software vulnerabilities.
- All guest OSes are located in the same local network.

Based on these assumptions, attackers can launch network attacks by spoofing the header of network packets.

4 Design Architecture

4.1 Architecture

Now, we present our design of the VNGS (Virtual Network Guard System). VNGS provides a lightweight packet filter built in virtual network interface. It can prevent malicious packets which is sent by attackers. Another feature is a centralized control center to deploy filtering program and monitor the virtual network interface on each host devicees. The entire VNGS is composed of two components as follow:

- Virtual Network Interface Filtering Controller
- Centralized Command and Control Center

Firstly, Virtual network interface filtering controller is a software emulating network interface control. It is used to collect raw network packets and run filtering process in virtual network interface control. If the filtering process intercepts the malicious packets, it will block the packets and inform the Center Command and Controller Center. Secondly, Centralized Command and Control Center is a management server with web interface. The administrators can use this server to monitor network environments in the data center, and the filtering process will send alert messages to control center when the attack is happened.

Virtual Network Interface Filtering Controller. The virtual machine monitors (i.e., QEMU, VirtualBox, etc.) usually equip software emulating network interface controller. We modify the control flow of the virtual network interface to add filtering mechanisms. Our design is based on the architecture of NE2000 network card emulator and Virtio network device. Fig 1 illustrates the architecture of NE2000 network card. Based on this method, Virtual Network Interface Filtering Controller is a modified software emulating network interface controller. We extract raw packets inside the network interface. Since the network interface is a hardware device in normal computer, the device driver processes the network packets as a number of instructions. The network device driver handles communication between hardware and software. In this case, the network interface receives instructions from network driver.

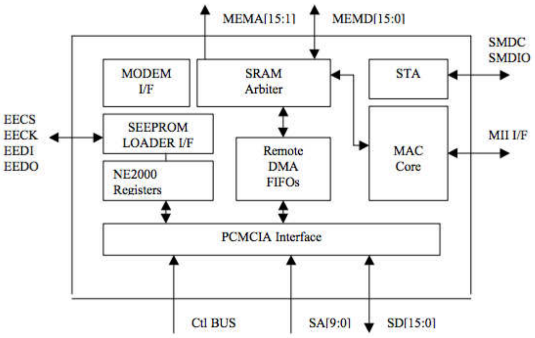


Fig. 1. Architecture of NE2000 network card

Fig 2 is the architecture of Virtual Network Interface Filtering Controller. In our design, we extract the packets from data transmit step and pass it to filtering process component. Filtering process is a program deployed for Centralized Command and Control Center. The input data of this process is raw packet and the output is a command to control network interface. Finally, network interface writes the packets to remote DMA according the result of filtering process.

Centralized Command and Control Center. Another component in our system is a centralized server which provides management console for administrators with web interface. It is a stand-alone server located in cloud environment. There are two main features of this server, collects malicious alert from virtual network interface and deploys filtering process to virtual network interface. Fig 3 illustrates the architecture of Centralized Command and Control Center.

The filtering process in the Virtual Network Interface Filtering Control will send messages back to Centralized Command and Control Center when it intercepts malicious packets. The communication between the host and server can be encrypted, so that the attackers can not monitor or counterfeit messages. The administrators can browse malicious logs by the web interface. Another feature is to filtering process deployment. In our design, the filtering process could be injected at run-time. After uploading a filtering process, the Deploy System will deliver the program to all virtual machines in the cloud environment.

4.2 Filtering Scheme

Now, we introduce a basic scheme to filter malicious ARP massages as an example. Based on the speciality of virtual environment, we can filter the packets in hardware level. Generally, the filtering mechanisms are usually implemented as a device driver in modern operating system such as Windows Filtering Platform[2]. However, the situation is different to the cloud environment. There are several reasons as follows. Firstly, each virtual machines in data center can be installed

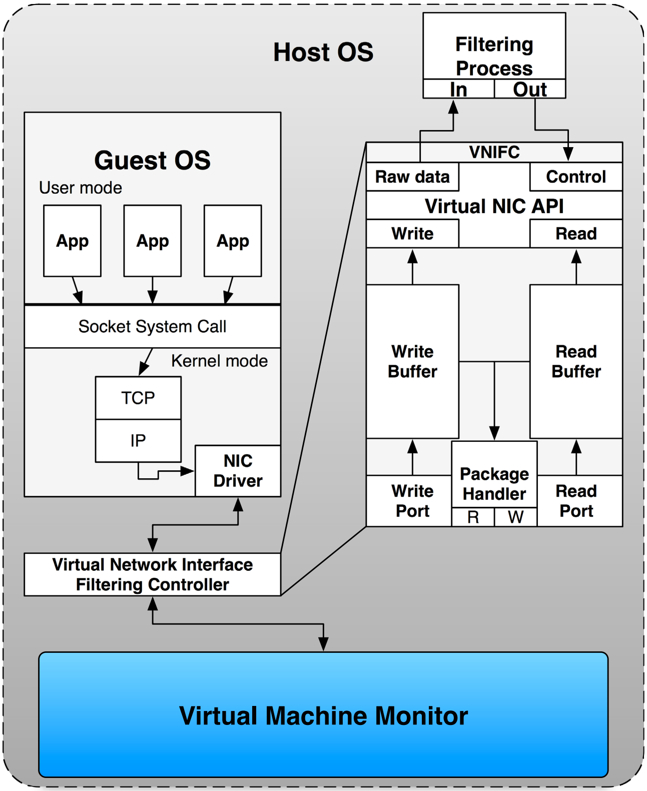


Fig. 2. Architecture of Virtual Network Interface Filtering Controller

with different operating systems. If the filtering mechanism is built in guest OS, we have to customize the driver in each operating systems. This needs higher costs and is inefficient. Secondly, even if we have installed modified guest OS, the attackers can bypass this mechanism. According to our assumptions, attackers have the ability to occupy the guest OSes and send modified instructions to the host. For these reasons, we build the filtering mechanisms in hardware level.

To protect each nodes in data center without ARP spoofing attack. We use a simple scheme to protect the local network. The first step of ARP spoofing attack is to fake ARP reply message. By faking reply message, the attack can poison ARP table of the switch. However, the MAC address of each network interface is hard coded on the device. In cloud environments, virtual machine monitor will configure the MAC when the network device is booted. It can not be changed during the device is used. According to this speciality, we inspect each network packets through the virtual network interface. If the source MAC address of the packet does not match with the MAC address hard coded on network interface, the filter process will drop this packet and inform administrator. Fig 4 show the flow chart of the filtering process our implementation.

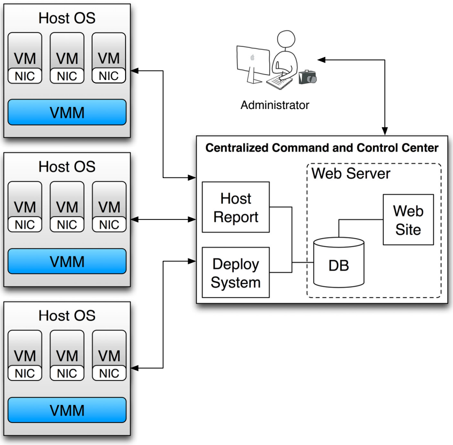


Fig. 3. Architecture of Centralized Command and Control Center

5 Performance Analysis

In our experiments, we execute the Centralized Command and Control Center on a Linux server with 2.0 GHz AMD 64 X2 Dual Core Processor. The virtual machine monitor is executed on another Linux server with 2.8 GHz Intel Xeon Processor. We used some experiments to test the capability of our system for surviving under ARP spoofing attack and provide the experimental results to illustrate the performance of our proposed system.

To evaluate the performance of virtual network, we measured our system on an physical machine hosting two VMs. We used an Intel Xeon W3530 2.8GHz with 8MB of L2 cache and 4 GB of main memory for our experiments. The host and guest OSes were running with Linux kernel 3.0.0-12. We set the main memory of the guest OSes to 256 MB. There were two kinds of network device model NE2000 and Virtio we used in our experiments. Both of them were modified to conform with our system. We loaded the filtering process used to defend ARP spoofing attacks. It dropped the packets which source MAC address did not match with the MAC address registered at network device, and filtered all the ARP reply packets which was not sent from correct host. The follows was the detail configurations:

- **NE2000.** A full virtualization virtual network device emulator often used in QEMU virtual machine. The maximum transfer rate of this device is 10 Megabits per second.
- **Virtio.** A paravirtualization virtual network device. By installing a special device driver to guest OS. It provides guest OSes direct access VMM without emulating physical network device. The maximum transfer rate of Vritio device is 1000 Megabits per second.

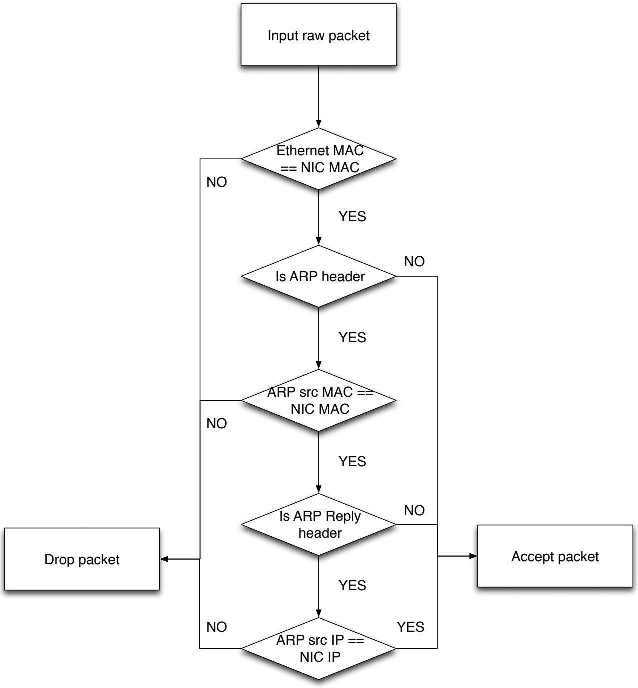


Fig. 4. Flow chart of our filtering process

In our experiments, we used latency and throughput to evaluate the performance. We measured round-trip latency by sending ICMP packets with 800 bytes packet size. The throughput was measured by using iperf [1] tool. Iperf was developed by NLANR/DAST as a modern alternative for measuring maximum TCP and UDP bandwidth performance. We measured the maximum TCP and UDP throughput of both NE2000 and Virtio network device models. The iperf was configured to send as fast as possible over 60 seconds with packet size 1500 Bytes (a standard MTU).

Round-trip Latency. We compared the latency and throughput of VNGS and normal QEMU network device here. The latencies were the average of 10 measurements which were shown in Fig 5. In the results of our experiments, the latency of both NE2000 and Virtio NIC increased overhead no more than 10%. We can see the overhead of our system did not increase too many costs. However, it might be due to our defend scheme did not focus on input packets. Therefore, the results appeared the performance of NE2000 is less than Virtio. VMM has to cost more time to handle network packets with full virtualization device.

Maximum Throughput. The Fig 6a and Fig 6b showed the maximum throughput of sending TCP and UDP packets. The results show NE2000 NIC decreased the throughput less than 3%. In our experiments, we turned off the Nagle's

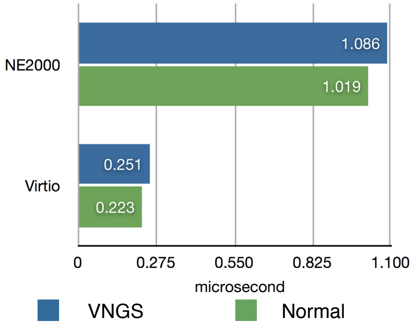
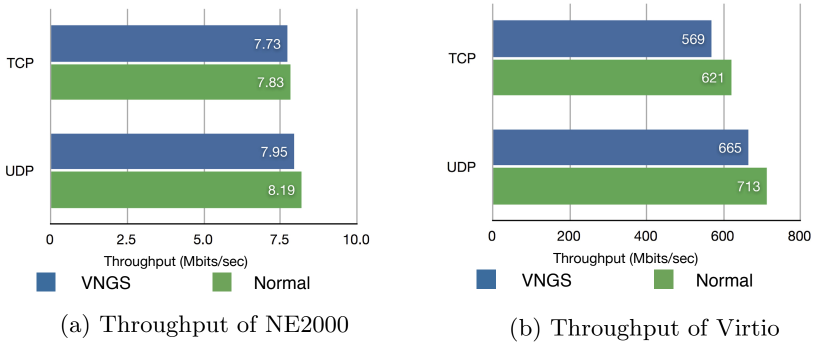


Fig. 5. Latency of the NE2000 and Virtio network device model



(a) Throughput of NE2000

(b) Throughput of Virtio

Fig. 6. VNGS performance test

algorithm [6] when the TCP packets were sent. The results showed that the throughput of NE2000 was almost no effect compared with normal QEMU network device. We considered that the VMM wastes lots of computing resource to emulate physical mechanism. Hence the overhead of filtering process was not significant.

On the other hand, the throughput of virtio device reduced almost 10%. The packets sending process had to suspend until it passed the filtering process. The Virtio device is different from full virtualization device such like NE2000, it does not need to emulate hardware circuit. The packets are moved from guest OS driver to the memory in VMM. There is why that the performance of virtio was better than NE2000 in our experiments.

6 Conclusion

In this paper, we provide and implement a system which is named VNGS against network security problem in cloud computing environments. With this system, administrators can use a centralize management server to monitor each VM in cloud computing environments. The filtering process build in virtual NIC can protect the local network without suffering from spoofed packets. We also provide a simple scheme against ARP spoofing attack. In our research, there are some ways to split the local network as many of VLANs. However, these schemes rely on some special hardware devices or the performance is inefficient. In our experiments, the results show that the overhead of our system is low, and the performance is acceptable compared to the other works.

Acknowledgment. This work was supported in part by the National Science Council, Taiwan, under Contracts NSC 100-2628-E-007-018-MY3 and NSC 103-2221-E-007-073-MY3.

References

1. Iperf: a network bandwidth testing tool, http://iperf.sourceforge.net/
2. Microsofat windows filtering platform, http://msdn.microsoft.com/en-us/windows/hardware/gg463267
3. Garfinkel, T., Pfaff, B., Chow, J., Rosenblum, M., Boneh, D.: Terra: A virtual machine-based platform for trusted computing. In: Proceedings of the 19th Symposium on Operating System Principles (SOSP 2003) (October 2003)
4. Garfinkel, T., Rosenblum, M.: A virtual machine introspection based architecture for intrusion detection. In: Proc. Network and Distributed Systems Security Symposium (February 2003)
5. Keller, E., Szefer, J., Rexford, J., Lee, R.B.: Nohype: virtualized cloud infrastructure without the virtualization. SIGARCH Comput. Archit. News 38(3), 350–361 (2010), http://doi.acm.org/10.1145/1816038.1816010
6. Nagle, J.: RFC 896: Congestion control in IP/TCP internetworks (January 1984), ftp://ftp.internic.net/rfc/rfc896.txt, ftp://ftp.math.utah.edu/pub/rfc/rfc896.txt, status: UNKNOWN
7. Ormandy, T.: An empirical study into the security exposure to hosts of hostile virtualized environments. Test, 1–10 (2007)
8. Pfaff, B., Pettit, J., Koponen, T., Amidon, K., Casado, M., Shenker, S.: Extending networking into the virtualization layer. In: Proc. of Workshop on Hot Topics in Networks (HotNets-VIII) (2009)
9. Szefer, J., Keller, E., Lee, R.B., Rexford, J.: Eliminating the hypervisor attack surface for a more secure cloud. In: Proceedings of the 18th ACM Conference on Computer and Communications Security, CCS 2011, pp. 401–412. ACM, New York (2011), http://doi.acm.org/10.1145/2046707.2046754

Towards SQL Injection Attacks Detection Mechanism Using Parse Tree

Tsu-Yang Wu[1,2], Jeng-Shyang Pan[1,2], Chien-Ming Chen[1,2],
and Chun-Wei Lin[1,2]

[1] Shenzhen Graduate School, Harbin Institute of Technology,
Shenzhen, 518055, China
[2] Shenzhen Key Laboratory of Internet Information Collaboration,
Shenzhen, 518055, China
{wutsuyang,jengshyangpan,chienming.taiwan}@gmail.com,
jerrylin@ieee.org

Abstract. With the development of network technology, database-driven web applications (apps) provide flexible, convenient, available, and various services for users. User can send requests to these web apps by using browser over the Internet to get services such as e-commerce services, entertainments, and financial services. Though web environments have several advantages, various security threats have been described. Among these threats, SQL injection attack (SQLIA) is one of the most serious threats. SQLIA is a code injection attack that exploits secure vulnerabilities consisting in source codes to attack databases. SQLIA allows attackers to bypass authentication, access private information, modify data, and even destroy databases. Since many sensitive and confidential data stored in database must be kept private and secure, a mechanism to detect SQLIAs for web environments is necessary. In this paper, we define a framework named DSD (Dynamic SQLIAs Detection) to counter SQLIAs in web environments. Then, a concrete detection mechanism based on DSD is proposed to detect SQLIAs by using parse tree. The experimental results are demonstrated that our mechanism has higher accuracy, lower false positive rate, and false negative rate.

Keywords: SQL injection attacks, parse tree, detection, web environments.

1 Introduction

With the development of network technology, web app can be accessed by using browser over the Internet. Database-driven web apps can supply more flexible, convenient, available, and various services for users and they have become the most important business model for companies in various fields. Especially, user can send requests to these web apps for getting services such as e-commerce services, entertainments, and financial services. In the past, some secure mechanisms for the Internet and database [6,11,17,23,24,28,29] had been published.

H. Sun et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 329, DOI: 10.1007/978-3-319-12286-1_38

Though web environments have several advantages, various security threats have been described. Among these threats, SQL injection attack (SQLIA) is one of the most serious threats for web apps [7,8,9,12]. An SQLIA which is an attacker inserts new SQL keywords or operators into an SQL query. With the altered query, an attacker can bypass authentication, obtain privacy information of users, modify or even destroy database.

Up to now, various methods to address SQLIAs for web applications have been proposed. Static analysis methods [4,10,21,25] attempt to prevent SQLIAs by finding all vulnerabilities before applications are deployed. These methods have no run time overhead. However, they analyze the source code of applications. It means that this kind of methods have two constraints. First, these methods are very host-language-specific; therefore, they cannot detect all kinds of SQLIAs. Second, these methods require to access source codes.

The mechanism Sania [20] detects SQLIAs in web applications during the development and debugging phases. The concepts of Sania are to capture an SQL query and generate an attack based on the syntax of potentially vulnerable spots in the captured SQL query. Then Sania can determine whether this query contains SQLIA flaws by comparing the parse trees of the intended SQL query.

Static and dynamic analysis mechanisms [13,14,22] have two phases, static analysis phase and dynamic analysis phase. It compares generated queries with normally expected queries. In the static phase, it analyzes a web applications source code to build models of legitimate queries. In the dynamic phase, queries are intercepted at run time and checked for conformity to the expected queries. Queries that do not match are rejected. Since this method requires to access source codes, it has restrictions when the source codes cannot be achieved. Moreover, the performance of this method depends on the quality of models built in static analysis phase.

Taint tracking mechanisms [2,3,5,16,26] attempt to solve the problem of SQLIAs by tracking user input and verifying that the input does not modify queries. However, this kind of method normally has additional requirements. The research [3,5]need to rewrite source codes to provide SQLIA detection. Other methods [2,16] require extra libraries to implement their design.

Several researchers [18,19,27] utilize machine learning technologies to detect SQLIAs. A detection method based on machine learning normally has two phases, learning phase and classification phase. In the learning phase, it utilizes a training set to build detection models. In classification phase, it judges if the query is an SQLIA with the models. The quality of training set will influence the performance of these methods.

Although several countermeasures of SQLIAs for web apps have been discussed, it still exist some drawbacks such as rewriting source codes of applications. In this paper, we first define a framework named DSD (Dynamic SQLIAs detection) to counter SQLIAs in web environments. Based on DSD, we propose an SQLIAs detecting mechanism by using parse tree. The main advantage of proposed mechanism is that it doesnt require to access the apps source code. Besides, DSD can be directly and easily embedded to existing web environments. Experimental results are demonstrated that our mechanism has higher accuracy rate, lower false positive rate, and lower false negative rate when detecting SQLIAs.

The rest of this paper is organized as follows. A framework DSD and a concrete detection mechanism are proposed in Section 2. In Section 3, we demonstrate the experimental results and conclusions are drawn in Section 4.

2 The Proposed Mechanism

In this section, we first define a framework named DSD (Dynamic SQLIAs detection) to counter SQLIAs in web environments. Based on this framework, we propose an SQLIAs detection mechanism. Notations used in this section are listed in the following:

- $G_r(\cdot)$: A function used for getting run-time stack, $G_r(\cdot) : q \rightarrow G_r(q)$, where $G_r(q)$, where q is a query.
- $G_t(\cdot)$: A function used for getting parse tree, $G_t(\cdot) : q \rightarrow G_t(q)$.
- H_c : A one-way hash function, $H_c : \{0,1\}^* \rightarrow \{0,1\}^k$.
- H_{id} : A one-way hash function, $H_{id} : \{0,1\}^* \rightarrow \{0,1,\ldots,n-1\}$, where n is an integer and depends on the number of slots in repository.
- C_t : A compressed parse tree, $C_t = H_c(G_t(q))$.
- C_r : A compressed run-time stack, $C_r = H_c(G_r(q))$.

2.1 DSD

The DSD consists of five units: Collector$_1$, Collector$_2$, Repository$_1$, Repository$_2$, and SQLIAs Agent and is depicted in Fig. 3.

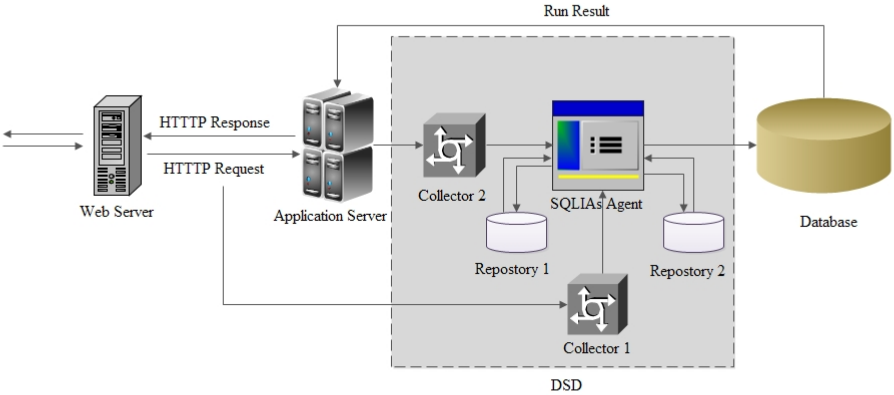


Fig. 1. The proposed framework DSD

DSD deployed between app server S and database is responsible to detect SQLIAs for a web app. The rough detecting process in DSD is described as followings.

1. User sends an HTTP request to a web application (app) within the server S. Then, the app dynamically generates a query q for this request. In this moment, Collector$_1$ captures HTTP request and sends it to SQLIAs Agent.
2. S sends q to DSD. Collector$_2$ captures some information which contains q and its run-time stack. Then, SQLIAs Agent interacts with Repository$_1$ and Repository$_2$ to verify q whether it is an SQLIA or not. If q is identified as an SQLIA, the agent will discard it. Otherwise, the agent will sent the query to the database.
3. The database executes q and sends back the result to the app.
4. The app generates a response to the user according to the result.

Note that Repository$_1$ and Repository$_2$ are two repositories. The structure of two repositories is depicted in Fig. 4. It contains a hash function H_{id} and an array A, where H_{id} is used to map a key key to the $H_{id}(key)^{th}$ slot of A. There are two operations called insertion and retrieve in the two repositories. In Repository$_1$, the insertion operation inserts 1 into $H_{id}(address)^{th}$ slot of A, i.e. $key = address$ and $value = 1$, and the retrieve operation gets 1 from the $H_{id}(address)^{th}$ slot of A. In Repository$_2$, the insertion operation inserts C_t into $H_{id}(C_t||C_r)^{th}$ slot of A, i.e. $key = (C_t||C_r)$ and $value = C_t$, and the retrieve operation gets C_t from the $H_{id}(C_t||C_r)^{th}$ slot of A.

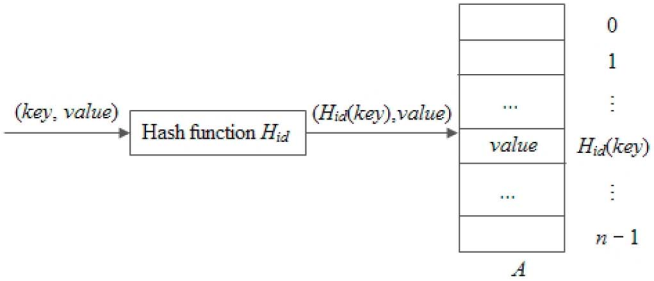


Fig. 2. The structure of repository

2.2 An Concrete SQLIAs Detection Mechanism

Based on DSD, we propose an SQLIAs detection mechanism. Our mechanism consists of two phases, classification and detection phases. When a user sends an HTTP request to an app, the classification phase is involved to identify the request whether it is first time access or non-first time access. After that, the detection phase provides SQLIA detection for this app in the above two cases.

Classification Phase. When a user sends an HTTP request to an app, Collector$_1$ captures and sends this request to SQLIAs Agent. Then, the app generates a query q corresponding to the request and sends q to Collector$_2$. It computes the corresponding run-time stack $G_r(q)$ of q and then sends q (original query) and $G_r(q)$ to

SQLIAs Agent. Upon receiving the information from Collector_1 and Collector_2, SQLIAs Agent obtains corresponding address and parameters from the request. Meanwhile, the agent parses the query q into a parse tree $G_t(q)$ and compresses it into a compressed parse tree $C_t = H_c(G_t(q))$. Then, SQLIAs Agent computes an index value $index = H_{id}(address)$ and then retrieves $A[index]$ in Repository_1. If the result equals to 1, the agent identifies the request as non-first time access and then invokes the non-first time access (NFTA) algorithm in the detection phase. Otherwise, the agent identifies the request as first time access and then invokes the first time access (FTA) algorithm in the detection phase. The detailed procedures of the classification phase are listed as follows.

1. Get address and parameters from an HTTP request. Note that address and parameters can be obtained in servlet program.
2. Get run-time stack $G_r(q)$ for query q. Note that the run-time stack can be implemented using high level languages such as Java.
3. Get parse tree $G_t(q)$ for query q. A parse tree can be implemented with open source tools.
4. Compute $C_t \leftarrow H_c(G_t(q))$.
5. Compute an index value $index$, where $index \leftarrow H_{id}(address)$.
6. Retrieve $value \leftarrow A[index]$ in Repository_1.
7. Compare $value$ with 1. If $value$ equals to 1, the NFTA algorithm is invoked. Otherwise, the FTA algorithm is invoked.

Detection Phase. In this phase, there are two cases which are the first time access and the non-first time access. The detail descriptions of the two cases are proposed as follows.

[First Time Access]

When an HTTP request is identified as first time access, SQLIAs agent inserts 1 into $H_{id}(address)^{th}$ slot of A in Repository_1. Meanwhile, the agent replaces all parameters of the query q with valid string such as "valid value" and obtains a transformed query q'. Note that q' is abstract valid and cannot be led to SQLIA. Then, SQLIAs agent parses the new query q' into a parse tree $G_t(q')$ and compresses it into a compressed parse tree $C'_t = H_c(G_t(q'))$. The agent compares the two compressed parse trees C_t with C'_t. If the both trees are equal, it means that the original query q is valid. In other words, the HTTP request is not an SQLIA. The original query is sent to the database. In this moment, SQLIAs agent compresses the run-time stack $G_r(q)$ of q and inserts the parse tree C_t into $H_{id}(C_t||C_r)^{th}$ slot of A in Repository_2. Otherwise, SQLIAs agent identifies q as SQLIAs and records it. The details procedures of FTA algorithm are listed as follows.

1. Insert 1 into slot $A[H_{id}(address)]$ in the Repository_1.
2. Get query q' by removing parameters from q.
3. Get parse tree $G_t(q')$ for query q'.
4. Compute $C'_t \leftarrow H_c(G_t(q'))$.
5. Compare with C_t with C'_t.

6. If $C_t \oplus C'_t$ equals to 0
   (a) q is a valid query.
   (b) Compute $C_r \leftarrow H_c(G_r(q))$.
   (c) Compute an index value $index$, where $index \leftarrow H_{id}(C_t||C_r)$.
   (d) Insert C_t into slot $A[index]$ in Repository$_2$.
7. Otherwise, q is identified as an SQLIA.

[Non-first Time Access]
When an HTTP request is identified as non-first time access, SQLIAs agent compresses the run-time stack $G_r(q)$ for original query q, ie. $C_r = H_c(Gr(q))$. Then, the agent retrieves old record C'_t in Repository$_2$ and compares the two compressed parse trees C'_t with C_t. If the both trees are equal, it means that the original query q is valid. In other words, the HTTP request is not an SQLIA. The original query is sent to the database. Otherwise, SQLIAs agent executes the procedures 2 to 7 of the FTA algorithm. The details procedures of NFTA algorithm are listed as follows.

1. Compute $C_r \leftarrow H_c(G_r(q))$.
2. Compute an index value $index$, where $index \leftarrow H_{id}(C_t||C_r)$.
3. Retrieve $C'_t \leftarrow A[index]$ in Repository$_2$.
4. Compare C_t with C'_t.
5. If $C_t \oplus C'_t$ equals to 0, q is a valid query.
6. Otherwise
   (a) Get query q' by removing parameters from q.
   (b) Get parse tree $G_t(q')$ for query q'.
   (c) Compute $C'_t \leftarrow H_c(G_t(q'))$.
   (d) Compare with C_t with C'_t.
   (e) If $C_t \oplus C'_t$ equals to 0
      i. q is a valid query.
      ii. Compute an index value $index$, where $index \leftarrow H_{id}(C_t||C_r)$.
      iii. Insert C_t into slot $A[index]$ in Repository$_2$.
   (f) Otherwise, q is identified as an SQLIA.

3 Experimental Results

In this section, we propose the experimental results for our mechanism. The experiments run on two standard PCs, PC$_A$ and PC$_B$. Both PC$_A$ and PC$_B$ have same hardware, where the processor is Intel(R) Core(TM) i5-2400M with 3.10 GHz, the RAM is 8GB, the hard disk is 500 GB, and the operating system is Windows 7. They are in the same local area network and can access each other. The architecture of experimental environment is shown in Fig. 5. PC$_B$ is used to simulate user (client) who can send accesses to related applications. We deploy one Tomcat 6 as web server and application server and MySQL 5.5 as database into PC$_A$. Then, the test application is deployed in the application server and all components of DSD are deployed in PC$_B$.

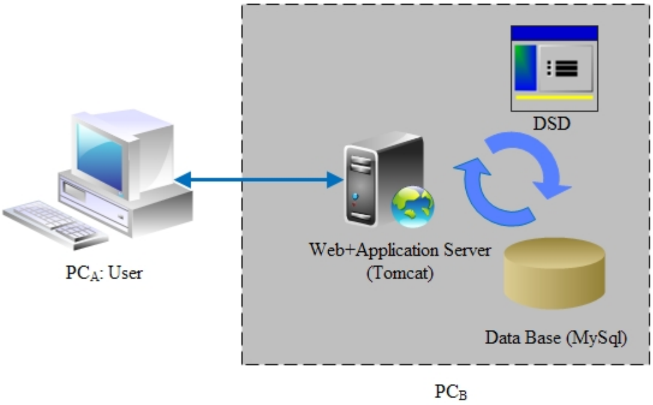


Fig. 3. Experimental environments

To demonstrate the precision and validity of our mechanism, we select four types of typical applications as test applications which are vulnerable to SQLIAs. The test applications consist of Employee Directory, Bookstore, Classifieds, and Portal [1]. Then, we choose test accesses from AMNESIA testbed suite [13,15] which is a static and dynamic analysis method to detect SQLIAs [14,15]. This suit contains a broad range of potential SQLIAs and legitimate accesses. However, it also contains some invalid accesses which cannot be accessed to database in our experimental environments. Hence, we need to remove these invalid accesses from this suit. After pre-processing the test applications, we have removed those invalid accesses and then obtain a test set which has 14674 test access including 13443 SQLIAs and 1231 legitimate accesses. The test set in our experiment is shown in Table 1.

Table 1. Test set

Types of applications	Legitimate Accesses	SQLIAs	Total
Employee Directory	124	3577	3701
Bookstore	124	3143	3267
Classifieds	348	3635	3983
Portal	635	3088	3723
Total	1231	13443	14674

Then, we demonstrate the detecting accuracy rate of our mechanism in Table 2. Obviously, the accuracy of our mechanism is over 99.9% for each type of typical application.

Finally, we show the false positive rate and the false negative rate of our mechanism in Tables 3 and 4. Obviously, the false positive rate of the proposed mechanism is less than 2% for each type of applications. The reason that false positives happened is that some queries cannot be parsed by DSD.

Table 2. The detecting accuracy of our mechanism

Types of applications	Total	Faults	Accuracy Rate
Employee Directory	3701	2	99.95%
Bookstore	3267	2	99.94%
Classifieds	3983	3	99.92%
Portal	3723	3	99.92%

Table 3. The false positive rate of our mechanism

Types of applications	Legitimate Accesses	False Positive	False Positive Rate
Employee Directory	122	2	1.64%
Bookstore	122	2	1.64%
Classifieds	348	3	0.86%
Portal	635	3	0.47%

Table 4. The false negative rate of our mechanism

Types of applications	SQLIAs	Successful Detections	False Negative Rate
Employee Directory	3577	3577	0%
Bookstore	3143	3143	0%
Classifieds	3635	3635	0%
Portal	3088	3088	0%

4 Conclusion

In this paper, we have proposed a framework DSD to counter SQLIAs for web environments. Based on this framework, a concrete detection mechanism has proposed to detect SQLIAs by using parse tree. Our mechanism does not require any access to source codes of apps. It means that DSD can be applied to existing web applications directly. Experimental results show that our mechanism has high accuracy rate, low false positive rate, and low time consumption. Hence, it is an efficient SQLIAs detection mechanism for web environments. In the future, we will compare our mechanism with previously proposed mechanisms.

Acknowledgments. This work is supported by Shenzhen Peacock Project of China (No. KQC201109020055A) and Shenzhen Strategic Emerging Industries Program of China (No. ZDSY20120613125016389).

References

1. `http://www.gotocode.com`
2. Bisht, P., Madhusudan, P., Venkatakrishnan, V.: Candid: Dynamic candidate evaluations for automatic prevention of sql injection attacks. ACM Transactions on Information and System Security (TISSEC) 13(2), 14 (2010)

3. Boyd, S.W., Keromytis, A.D.: Sqlrand: Preventing sql injection attacks. In: Jakobsson, M., Yung, M., Zhou, J. (eds.) ACNS 2004. LNCS, vol. 3089, pp. 292–302. Springer, Heidelberg (2004)
4. Bravenboer, M., Dolstra, E., Visser, E.: Preventing injection attacks with syntax embeddings. In: Proceedings of the 6th International Conference on Generative Programming and Component Engineering, pp. 3–12. ACM (2007)
5. Buehrer, G., Weide, B.W., Sivilotti, P.A.: Using parse tree validation to prevent sql injection attacks. In: Proceedings of the 5th International Workshop on Software Engineering and Middleware, pp. 106–113. ACM (2005)
6. Chen, C.M., Zheng, X., Wu, T.Y.: A complete hierarchical key management scheme for heterogeneous wireless sensor networks. The Scientific World Journal 2014, Article ID 816549, 13 pages (2014)
7. Christey, S., Martin, R.A.: Vulnerability type distributions in cve (2007)
8. Clarke, J.: SQL injection attacks and defense. Elsevier (2012)
9. Dhamankar, R., Dausin, M., Eisenbarth, M., King, J., Kandek, W., Ullrich, J., Skoudis, E., Lee, R.: The top cyber security risks. TippingPoint, Qualys, the Internet Storm Center and the SANS Institute faculty. Tech. Rep. (2009)
10. Fu, X., Lu, X., Peltsverger, B., Chen, S., Qian, K., Tao, L.: A static analysis framework for detecting sql injection vulnerabilities. In: Proceedings of the 31st Annual International Computer Software and Applications Conference (COMPSAC 2007), vol. 1, pp. 87–96. IEEE (2007)
11. Guo, C., Chang, C.C., Sun, C.Y.: Chaotic maps-based mutual authentication and key agreement using smart cards for wireless communications. Journal of Information Hiding and Multimedia Signal Processing 4(2), 99–109 (2013)
12. Halfond, W., Viegas, J., Orso, A.: A classification of sql-injection attacks and countermeasures. In: Proceedings of the IEEE International Symposium on Secure Software Engineering, Arlington, VA, USA, pp. 13–15 (2006)
13. Halfond, W.G., Orso, A.: Amnesia: analysis and monitoring for neutralizing sql-injection attacks. In: Proceedings of the 20th IEEE/ACM International Conference on Automated Software Engineering, pp. 174–183. ACM (2005)
14. Halfond, W.G., Orso, A.: Preventing sql injection attacks using amnesia. In: Proceedings of the 28th International Conference on Software Engineering, pp. 795–798. ACM (2006)
15. Halfond, W.G., Orso, A., Manolios, P.: Using positive tainting and syntax-aware evaluation to counter sql injection attacks. In: Proceedings of the 14th ACM SIGSOFT International Symposium on Foundations of Software Engineering, pp. 175–185. ACM (2006)
16. Halfond, W.G., Orso, A., Manolios, P.: Wasp: Protecting web applications using positive tainting and syntax-aware evaluation. IEEE Transactions on Software Engineering 34(1), 65–81 (2008)
17. He, B.Z., Chen, C.M., Su, Y.P., Sun, H.M.: A defence scheme against identity theft attack based on multiple social networks. Expert Systems with Applications 41(5), 2345–2352 (2014)
18. Huang, Y.W., Huang, S.K., Lin, T.P., Tsai, C.H.: Web application security assessment by fault injection and behavior monitoring. In: Proceedings of the 12th International Conference on World Wide Web, pp. 148–159. ACM (2003)
19. Komiya, R., Paik, I., Hisada, M.: Classification of malicious web code by machine learning. In: Proceedings of the 3rd International Conference on Awareness Science and Technology (iCAST 2011), pp. 406–411. IEEE (2011)

20. Kosuga, Y., Kernel, K., Hanaoka, M., Hishiyama, M., Takahama, Y.: Sania: Syntactic and semantic analysis for automated testing against sql injection. In: 23th Annual Computer Security Applications Conference (ACSAC 2007), pp. 107–117. IEEE (2007)
21. Lam, M.S., Whaley, J., Livshits, V.B., Martin, M.C., Avots, D., Carbin, M., Unkel, C.: Context-sensitive program analysis as database queries. In: Proceedings of the 24th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, pp. 1–12. ACM (2005)
22. Lee, I., Jeong, S., Yeo, S., Moon, J.: A novel method for sql injection attack detection based on removing sql query attribute values. Mathematical and Computer Modelling 55(1), 58–68 (2012)
23. Lin, C.W., Hong, T.P., Chang, C.C., Wang, S.L.: A greedy-based approach for hiding sensitive itemsets by transaction insertion. Journal of Information Hiding and Multimedia Signal Processing 4(4), 201–227 (2013)
24. Lin, C.W., Hong, T.P., Hsu, H.C.: Reducing side effects of hiding sensitive itemsets in privacy preserving data mining. The Scientific World Journal 2014, Article ID 235837, 12 pages (2014)
25. McClure, R.A., Kruger, I.H.: Sql dom: compile time checking of dynamic sql statements. In: Proceedings of the 27th International Conference on Software Engineering (ICSE 2005), pp. 88–96. IEEE (2005)
26. Mitropoulos, D., Spinellis, D.: Sdriver: Location-specific signatures prevent sql injection attacks. Computers & Security 28(3), 121–129 (2009)
27. Valeur, F., Mutz, D., Vigna, G.: A learning-based approach to the detection of sql attacks. In: Julisch, K., Kruegel, C. (eds.) DIMVA 2005. LNCS, vol. 3548, pp. 123–140. Springer, Heidelberg (2005)
28. Wu, T.Y., Tsai, T.T., Tseng, Y.M.: A revocable id-based signcryption scheme. Journal of Information Hiding and Multimedia Signal Processing 3(3), 240–251 (2012)
29. Wu, T.Y., Tsai, T.T., Tseng, Y.M.: A provably secure revocable id-based authenticated group key exchange protocol with identifying malicious participants. The Scientific World Journal 2014, Article ID 367264, 10 pages (2014)

No-Reference Image Quality Assessment in Spatial Domain

Tao Sun[1], Xingjie Zhu[2,*], Jeng-Shyang Pan[3], Jiajun Wen[3], and Fanqiang Meng[2]

[1] Zhangjiakou Power Supply Company, Zhangjiakou City, Hebei Province China
[2] Research and Development Center, Beijing HuaRong JingDun Technology Co., Ltd., Beijing, China
[3] Shenzhen Graduate School, Harbin Institute of Technology, 518055, Shenzhen, China
zhu.xingjie322@gmail.com

Abstract. With the development of computer vision, there has been an increasing need to develop objective quality measurement techniques that can predict image quality automatically. In this paper, we present a complex No-reference image quality assessment (NR IQA) algorithm, which mainly consists of two steps. The first step uses Gabor filters to obtain the feature images with different frequencies and orientations, so as to extract the energy and entropy features of each sub-image. The second step uses the Linear least squares to obtain the parameters for IQA. We conduct experiments in LIVE IQA Database to verify our method. The experimental results show that the proposed method is much more competitive than other state of the art Full-reference (FR) or NR algorithms.

Keywords: Gabor Filter, Least Squares, Image Quality, Energy, Entropy.

1 Introduction

For the rapid and massive dissemination of the digital images and videos it is essential to provide a required level of customer satisfaction [1, 2]. Since a large quantity of images are of low quality, it is important to evaluate image quality automatically. IQA can be applied in many fields. For example, they can be used to optimize algorithms and parameter settings for image processing systems. They can also be embedded into video monitoring systems to evaluate image quality. The most widely used objective image quality metrics are Peak Signal-to-Nose Ratio (PSNR) [3] and blind image quality indices (BIQI) [4]. However, they are not in accord with perceived quality measurement.

In the literature [5-8] , the majority of proposed IQA methods require a reference image with high quality image. In this paper, we propose a new framework that can easily assess the quality of distorted images without any reference image. Generally, designing objective NR quality measurement algorithms is a very crucial issue. This is mainly due to the limited understanding of the human visual system (HVS) for us. Although only a few methods have been proposed in the literatures [9-11] for objective NR IQA, this topic has attracted a great deal of attention recently.

[*] Corresponding author.

H. Sun et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 329, DOI: 10.1007/978-3-319-12286-1_39

Moorthy et al. [12] proposed an Distortion Identification-based Image Verity and Integrity Evaluation (DIIVINE) method that assesses the quality of a distorted image without a reference image. This approach assumes that natural scenes possess certain statistical properties which are altered in the presence of distortion. Liu et al. [13] studied the efficacy of utilizing the curvelet transform to learn a NR IQA model. In this paper, a set of statistical features are extracted from the image through the curvelet representation. The extracted features include the coordinates of the maxima of the log-histograms of the curvelet coefficients values, and the energy distributions of both orientation and scale in the curvelet domain. Although above mentioned algorithms deliver prominent performances, there still remains huge space for performance improvement in the creation of NR IQA models. In our paper, we devote to design an effective and efficient IQA model based on the orientation and the frequency distribution of an image. Experimental results show that a set of energy and entropy features extracted in the spatial domain are highly relevant to natural image quality.

The rest of the paper is organized as follows. In Section2, we describe the detailed of feature extraction from images, and introduce a method to obtain the optimal parameters. We evaluate the performance of the proposed algorithm in Section3 and conclude the paper in Section 4.

2 The Proposed Method

We present the flow chart of the proposed no-reference IQA method in the figure 1. The proposed IQA method mainly consists of two steps. The first step uses Gabor filters to obtain the feature images in different frequencies and orientations, then, extracts the energy and entropy features of each sub-image. The second step uses the Linear least squares to obtain the parameters for IQA.

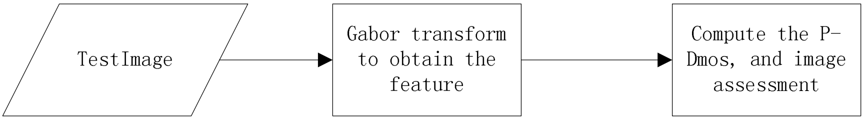

TestImage
Gabor transform to obtain the feature
Compute the P-Dmos, and image assessment

Fig. 1. Flow chart of our method

2.1 The First Step: Gabor Feature Extraction

As we all know that, frequency and orientation representations of Gabor filters are in accord with the human visual perceptions. Gabor filters have been found to be particularly appropriate for texture representation and discrimination [14]. In our paper, we use a set of Gabor filters with different frequencies and orientations to extract useful features from a test image to obtain the quality scores.

First, we transform the color image into grayscale image according to the following function:

$$Gray = R*0.299 + G*0.587 + B*0.114 \tag{1}$$

Then we use Gabor filters to obtain the filtered images. Gabor filters are of various types through tuning of the parameters in terms of frequency and orientation. By varying these parameters, a filter is able to pass any elliptical region of spatial frequencies. The Gabor function g(x, y) is as follows:

$$g(x,y)=\left(\frac{1}{2\pi\sigma_x\sigma_y}\right)\exp\left[-\frac{1}{2}\left(\frac{\vec{x}^2}{\sigma_x^2}+\frac{\vec{y}^2}{\sigma_y^2}\right)+2\pi jW\vec{x}\right] \tag{2}$$

$$\vec{x} = x\cos\theta + y\sin\theta \tag{3}$$

$$\vec{y} = -x\sin\theta + y\sin\theta \tag{4}$$

where σ_x and σ_y are the scaling parameters of the filter and determine the effective size of neighborhood of pixel. $\theta(0,\pi)$ means the orientation of the Gabor filters. W is the radial frequency of the sinusoid. The Fourier transform of the Gabor function in (2) is given by

$$G(u,v)=\exp\left[-\frac{1}{2}\left(\frac{(u-W)^2}{\delta_u^2}+\frac{v^2}{\delta_v^2}\right)\right] \tag{5}$$

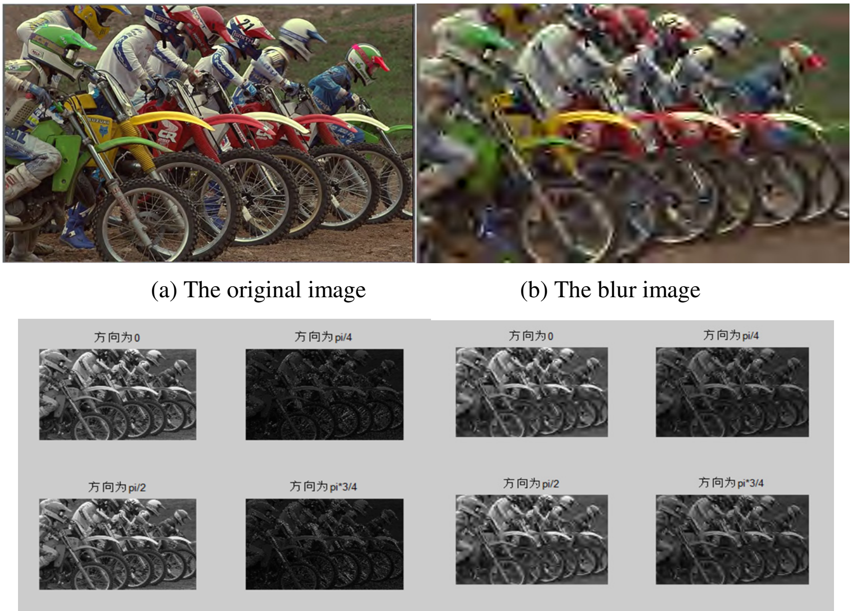
where $\sigma_u = 1/2\pi\sigma_x$, σv $= 1/2\pi\sigma_y$.The Fourier domain representation in (5) specifies the amount by which the filter modifies each frequency component of the input image.

In our paper, we use four different orientations and two different frequencies to obtain the filtered images. Figure 2 shows the Gabor filters image with different parameters setting in frequency domain and orientation. According to the filtered image, we can compute the energy and entropy of the filtered image as follows:

$$\text{energy}:\ f_1 = \sum_{i=0}^{L-1}\sum_{j=0}^{L-1} p_\delta^2(i,j) \tag{6}$$

$$\text{entropy}:\ f_2 = -\sum_{i=0}^{L-1}\sum_{j=0}^{L-1} p_\delta(i,j)\lg p_\delta(i,j) \tag{7}$$

where $p_\delta(i,j) = f(i,j)/N^2$, and we can know that i and j mean the pixel value and neighbourhood value respectively; f(i, j) is the frequency at location (i, j) and N is the scales of the image respectively. Then, we can compute the feature vector $\vec{p}$ that has 16 elements, $\vec{p}$ =(e1, e2, e3, e4, e5, e6, e7, e8, n1, n2, n3, n4, n5, n6, n7, n8), where e1,,,,, e7 stand for the values of the energy features and n1,,,,, n7 stand for the values of the entropy features of the 8 filter images, respectively.



(a) The original image (b) The blur image

(c) the Gabor filter image of (a) (d) the Gabor filter image of (b)

Fig. 2. The result of Gabor filter in four orientations of 0,45,90,135, combined with 0.3 frequency

2.2 The Second Step: Parameter Training

According to the last section, we can obtain the feature vector $\vec{p}$ of the test image. In this section, we mainly talk about how to obtain the weight vector $\vec{q}$ =(x1, x2, x3, x4, x5, x6, x7, x8, y1, y2, y3, y4, y5, y6, y7, y8) (where x1,,,,, x8 stand for the weights of the energy features and y1,,,,, y8 stand for the weights of the entropy features of 8 filter images, respectively) of this features to constitute a quality assessment model. As we all know that Linear least squares regression is by far the most widely used modeling method. Not only is linear least squares regression the most widely used modeling method, but it has been adapted to a broad range of situations that are outside its direct scope. Then, we use linear least squares to obtain $\vec{q}$, and the final objective quality prediction can be write as follows:

$$P_{Dmos} = \vec{q}^{T}\vec{p} = x_{1*}e_1 + x_{2*}e_2 + x_{3*}e_3 + x_{4*}e_4 + x_{5*}e_5 + x_{6*}e_6 + x_{7*}e_7 + x_{8*}e_8 + \\ y_{1*}n_1 + y_{2\,*}n_2 + y_{3*}n_3 + y_{4*}n_4 + y_{5*}n_5 + y_{6*}n_6 + y_{7*}n_7 + y_{8*}n_8 \tag{8}$$

Where each explanatory variable in the function is multiplied by an unknown parameter, there is at most one unknown parameter with no corresponding explanatory variable, and all of the individual terms are summed to produce the final function value.

We get the associated human differential mean opinion scores (DMOS), which are representative of the perceived quality of the image, to train parameter vector $\vec{q}$. The compute function can rewrite as follows:

$$\mathrm{E} = \min \left| \sum_{n=1}^{\mathrm{N}} \left(\mathrm{P}_{\mathrm{Dmos}} - \mathrm{DOMS} \right)^2 \right| \tag{9}$$

where N is the number of training samples. when obtaining the weight vector $\vec{q}$ according to equation (9), we can evaluate the $\mathrm{P}_{\mathrm{Dmos}}$ of test sample rely on $\vec{q}$.

3 Experiments

We tested our method on the popular Live IQA database[15], which consists of 29 reference images and 779 distorted images. It also give DMOS which are representative of the perceived quality of the image. In our paper, we conducted experiments on JPEG and JPEG2000 compression Images to verify our method.

We iteratively divided the LIVE database into training sample and test sample. Using the training sets to train the regression models which are required in algorithm parameters. Then we run the algorithm on the test sets to evaluate its performance. We split the training and the test sets according to the following rules: (1) the training and test sets were absolutely separated by content; (2) we randomly selected the training and tests from the LIVE IQA database 10 times and evaluated the performance on each test set. The average performance over 10 times were used as the final performance evaluation. The performance evaluation indices include the Spear-man's Rank Ordered Correlation Coefficient (SROCC) and the linear correlation coefficient (LCC) between the predicted objective quality scores($\mathrm{P}_{\mathrm{Dmos}}$) and the given DMOS.

3.1 Correlation of Each Feature Vector with Human Perception

In order to validate the effectiveness of each feature vector, we utilize different features to conduct the performance evaluation on JPEG and JPEG2000 compression Images. Tables 1–3 show the median experimental results on our model over 10 iterations of random 80–20% train-test splits using different feature vector.

Table 1. Energy feature

	SROCC	LCC
JPEG	0.8549	0.8590
JPEG2000	0.8092	0.7949

Table 2. Entropy feature

	SROCC	LCC
JPEG	0.8723	0.8630
JPEG2000	0.8298	0.8373

Table 3. Combine energy and entropy feature

	SROCC	LCC
JPEG	0.9495	0.9435
JPEG2000	0.9237	0.9304

3.2 Performance Comparison with Other IQA Models

We also compared the performance of our method with PSNR (which is a FR approach) and three other NR approaches (BIQI [4], DIIVINE [12], CurveletQA[13]). In order to conduct a fair comparison, for the FR approach, we excluded all reference images from the LIVE database and used only the distorted images for testing. Since the performances of other three NR methods have been given in [13], we don't need to verify these performance again. Our algorithms' performance evaluations were obtained also by 80% training – 20% testing over 10 times.

Table 4. SROCC on the LIVE IQA database

	JPEG	JPEG2000
PSNR	0.8515	0.8837
BIQI[4]	0.8414	0.7603
DIIVINE[12]	0.8921	0.9352
CurveletQA[13]	0.9117	0.9376
Our method	0.9495	0.9237

Table 5. LCC on the LIVE IQA database

	JPEG	JPEG2000
PSNR	0.8515	0.8837
BIQI[4]	0.8414	0.7603
DIIVINE[12]	0.9097	0.9409
CurveletQA[13]	0.9281	0.9465
Our method	0.9435	0.9304

It can be concluded from the experimental results (Table 4-5) that our method correlates well with the human subjective opinions of image quality, which shows that the proposed method is statistically superior to the most popular IQA approaches such as PSNR, BIQI, CurveletQA , and DIIVINE in JPEG database. Although our method is inferior to the top-performing CurveletQA, and DIIVINE in JPEG2000 database, its performance is almost the same as these methods.

4 Conclusion

The proposed two-step no-reference IQA method has the following noticeable merits: first, we use Gabor filters to get the feature sub-image, and extract the energy and entropy feature vector from the sub-image. In our experiment, we found that frequency and orientation representations of Gabor filters are similar to those of the human visual system, and they have been found to be particularly appropriate for texture representation and discrimination. Second, we use Linear least squares to train the test parameters for assessment. Practically speaking, linear least squares regression makes good use of the data. Promising results can be obtained with relatively small data sets. Third, the proposed method can efficiently and precisely predict the quality of test image. The experimental results show that the proposed method performs very well on the LIVE IQA databases comparing with other methods.

References

1. Loukhaoukha, K.: On The Security of Digital Watermarking Scheme Based on Singular Value Decomposition and Tiny Genetic Algorithm. Journal of Information Hiding and Multimedia Signal Processing 3(2), 135–141 (2012)
2. AlQaheri, H., Mustafi, A., Banerjee, S.: Digital Watermarking using Ant Colony Optimization in Fractional Fourier Domain. Journal of Information Hiding and Multimedia Signal Processing 1(3), 179–189 (2010)
3. Thu, H., Ghanbari, M.: Scope of validity of PSNR in image/video quality assessment. Electronics Letters 44(13), 800, doi:10.1049/el:20080522
4. Moorthy, A.K., Bovik, A.C.: A two-step framework for constructing blind image quality indices. IEEE Signal Process. Lett. 17(5), 513–516 (2006)
5. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncell, E.P.: IQA: from error visibility to structural similarity. IEEE Trans. Image Process. 13(4), 600–612 (2004)
6. Sheikh, H.R., Bovik, A.C.: Image information and visual quality. IEEE Trans. Image Process. 2006 15(2), 430–444 (2006)
7. Chandler, D.M., Hemami, S.S.: VSNR: a wavelet-based visual signal-to-noise ratio for natural images. IEEE Trans. Image Process 2007 16(9), 2284–2298 (2007)
8. Sheikh, H.R., Bovik, A.C.: Image information and visual quality. IEEE Trans. Image Process. 15(2), 430–444 (2006)
9. Liu, L., Dong, H., Huang, H., Bovik, A.C.: No-reference IQA in curvelet domain. Signal Processing: Image Communication (February 2014)

10. Yu, F.-X., Lei, Y.-Q., Wang, Y.-G., Lu, Z.-M.: Robust Image Hashing Based on Statistical Invariance of DCT Coefficients. Journal of Information Hiding and Multimedia Signal Processing 1(4), 286–291 (2010)
11. Shen, J., Li, Q., Erlebacher, G.: Hybrid no-reference natural IQA of noisy, blurry, JPEG2000, and JPEG Images. IEEE Trans. Image Process. 20(9), 2089–2098 (2011)
12. Moorthy, A.K., Bovik, A.C.: Blind IQA: From natural scene statistics to perceptual quality. IEEE Trans. Image Process. 20(12), 3350–3364 (2011)
13. Liu, L., Dong, H., Huang, H., Bovik, A.C.: No-reference IQA in curvelet domain. Signal Processing: Image Communication 29, 494–505 (2014)
14. Yang, C.Y., Lin, C.H., Hu, W.C.: Reversible Data Hiding for High-Quality Images Based on Integer Wavelet Transform. Journal of Information Hiding and Mul-timedia Signal Processing 3(2), 142–150 (2012)
15. Sheikh, H.R., Wang, Z., Cormack, L., Bovik, A.C.: LIVE IQA Database Release 2, http://live.ece.utexas.edu/research/quality

MDPAS: Markov Decision Process Based Adaptive Security for Sensors in Internet of Things

Eric Ke Wang[1,*], Tsu-Yang Wu[1], Chien-Ming Chen[1], Yuming Ye[1], Zhujin Zhang[1], and Futai Zou[2]

[1] Shenzhen Key Laboratory of Internet Information Collaboration,
Shenzhen Graduate School, Harbin Institute of Technology, China
[2] School of Information Security Engineering, Shanghai Jiaotong University
{wk_hit,yym}@hitsz.edu.cn,
{wutsuyang,chienming.taiwan}@gmail.com, zoufutai@sjtu.edu.cn

Abstract. Nowadays chipped based sensors and RFID tags are widely employed in Internet of Things; however, for those devices, effective and flexible security mechanisms lack. In this paper we study the security requirement and propose an adaptive security framework for sensors in Internet of things, which provides dynamic confidentiality, authenticity and integrity in the networks with relative suitable overhead by context aware computing, decision making and dynamic enforcement of policies. We employ Markov Decision Process to make the decisions of security actions and adopt aspect-oriented programming technique to enforce the security policies dynamically in the working networks. We made simulations of our framework, and the performance is encouraging.

Keywords: Adaptive security, Markov Decision Process, Aspect-Oriented Programming.

1 Introduction

The Internet of Things (IoT), a highly heterogeneous network with various kinds of objects, has been a very hot area in academic and industrial world. IoT mainly includes three types of communication ways, human-to-human, thing-to-thing and human-to-thing. Since people become more and more familiar with IoT, security problems remain challenging it. Many IoT devices(such as sensors, RFID tags) may have limited resource and low power, while, higher secure measures always come with more resources and energy consumption since they cost a lot of computation and communication. In other words, security measures and power saving often stand in the opposite side with each other. On one hand, IoT designer are willing to employ high secure measures to increase their security, on the other hand, they hope those measures cost as less as possible in IOT.

Therefore, it is desirable to achieve security and low cost at the same time for sensors in IOT. Traditional security measures of sensor networks are mainly targeting on selecting or propose low power security algorithms or protocols. While, in our previous study, we found that traditional security mechanisms lack flexibility,

[*] Corresponding author.

H. Sun et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 329, DOI: 10.1007/978-3-319-12286-1_40

security measures for IoT should be more adaptive to current context. Therefore, in this paper we propose an adaptive security framework for IoT, which is called MDPAS. It mainly provides dynamic secure measures by context aware computing, decision making and dynamic enforcement of policies. In this paper, we explain how MDPAS can adaptively adjust the security policies to meet the various security requirement of complicated environment. Moreover, we have made simulation and the performance is invigorating.

The paper is organized as follows: section 1 is the introduction to introduce the background of trust issues in IOT; section 2 is the relative works about the trust models; in section 3 we introduce a security model for the security threats in IOT; in section 4 we propose the adaptive framework for sensors in IoT; in Section 5 we made the simulation and analysis; the final conclusion is in section 5.

2 Related Work

Adaptive security mechanism is that the security actions can be self-decided and self-executed in the dynamic systems. It mainly includes three parts, context awareness, security action decision making and policies enforcement. Context awareness and inference is relative mature field[1-5], however, currently most of context awareness computing is for the sufficient resources computing environment, while, for many IoT devices with low computation and communication resources, the support of context awareness computing is limited.

Besides, on the problems of security actions decision making and dynamic enforcement, Preda [6] proposed how to deploy the access control polices according to current situation in the limited resources devices, but it mainly solves the related problems of access control. Mardziel [7] proposed an enforcement way of security policies based on knowledge, but it mainly identify the different level of security policies according to limited knowledge and it needs the active interaction of users. Aloulou [8] employs formal method to define the security polices for mobile agent systems and propose a dynamic enforcement way, Rafailidis [9] proposed a reference monitor inlining approach that treat input injection vulnerabilities as a crosscutting concern. But it is mainly for web application.

Currently, adaptive security mechanism is employed in many resources sufficient systems, however, for limited resources devices in IoT, adaptive security mechanisms lack. Therefore, we propose an adaptive security framework for IoT to make IoT devices can dynamically decide the security actions and make enforcement according to situations.

3 Cyber Physical Adaptive Security Framework for IoT

IoT is a heterogeneity and complex system, and it may have various security requirements. However, if we enforce complete security measures to satisfy all the requirements, it is a great burden for IoT devices especially for limited resources devices. Therefore, as we observation, we propose an adaptive security framework for IoT that can adapt security measures to current context.

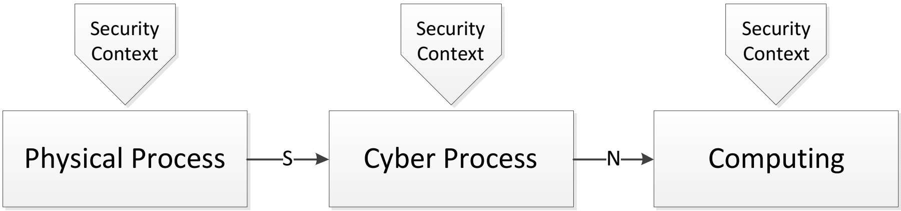


Fig. 1. Adaptive security framework for IoT

As Figure 1 shows, it is an adaptive security framework for IoT which incorporates various security measurements such as authentication, encryption, key agreement protocol, access control in sensing, networking, and computing processes. Commonly, IoT connects physical things and cyber world. The process flow can be divided into physical process, cyber process and computing process. Therefore, security mechanism for IoT devices can be dynamically adapting to context by the assistance of context coupling at each key process. We call this kind of security mechanism adaptive security framework. It mainly includes context aware computing, decision making process and dynamic enforcement.

3.1 Context Aware Computing

In our framework, we mainly tackle security-relevant context which consists of the set of contextual attributes that can be used to characterize the situation of an entity, whose value affects the choice of the most appropriate measures or the configuration of the network to protect information from unauthorized access, use, disclosure, disruption, modification or destruction, and based on it, we provide confidentiality, integrity and availability. When attacks occur, the attack model and the adversary types can also be one of the contextual attributes. The values of security-relevant contextual attributes affect the choice of the most appropriate controls because they impact the likelihood of certain threats to confidentiality, integrity, and availability being realized. Therefore, based on their values, the most appropriate controls and configuration of those controls can be employed to mitigate those threats.

We use C_i represent one context, then a multiple context set *MS* includes *k* types of context information which can be inference to a final context *obj_cs* for security policies (we design a context inference engine to make context aggregation and inference):

$$MS=\{C_1,C_2,\ldots\ldots,C_k\}\text{-->}obj\_cs \qquad (1)$$

3.2 Dynamic Markov Decision Process

Selecting security actions based on current context is a typical decision making process, so we adopt Markov Decision Process to solve the problem since it is one of the best decision process for random dynamic network. We can look at each security action as a state; the decision of each action is to select one of the best security actions. Then after *n* piece of decisions are made sequentially, the accumulate energy cost are expected to be low.

In each step, the decision making depends on the current situation, and with sequential decisions are made, the whole security policies is the optimal. Since the decision making steps are not infinite, so we adopt finite Markov decision process to solve it. The basic idea is: we use maximum of decision metric in the whole process as the goal of the decisions to build a Finite Markov Decision Process model, in order to find out the set of best security actions from candidates.

T Phases Markov Decision Process (as shown in Following Tetrad, Equation 2):

$$\{E; \quad D(i) \mid i \in E; \quad P_{ij}(\alpha) \mid i, j \in E, \alpha \in D(i); \quad r(i,a) \mid i \in E, \quad \alpha \in D(i)\} \tag{2}$$

They are states space, decision set, state transit probability and valued reward. They includes that **Set of States** E, **Set of Decisions** $D_T(i)$, **Probability of States Transit** $P_{ij}(\alpha)$, **Expected Reward** $r(i,\alpha)$, **Policy** θ;

The **Accumulation computation** $u_t^*(i)$:

$$u_t^*(i) = max_{d \in D_S}\{r_t(i,d_t(i)) + \sum_{j \in S} p_t(j \mid i, d_t(i)) u_{t+1}^*(j)\} \quad j = T-1, T-2, \ldots, t \tag{3}$$

Then we can find out the policies set $\pi = (d_1(i), d_2(i), \ldots., d_T(i))$, that is the final path result of decision.

3.3 Implementation Method of Dynamic Enforcement

After the policies of secure actions are decided, the actions should be integrated into real-time network system to realize the adaptive security measures. Thus the dynamic enforcement is required, as shown in figure 2, it has an objective that dynamically integrate security actions into main programs. When the method is called, firstly it adjudges if the method need security control, if yes, execute the decision making for deciding security policies and integrate them into programs. As our previous work, we adopt Aspect-Oriented Programming(AOP) [10]method to realize the dynamic enforcement of security policies.

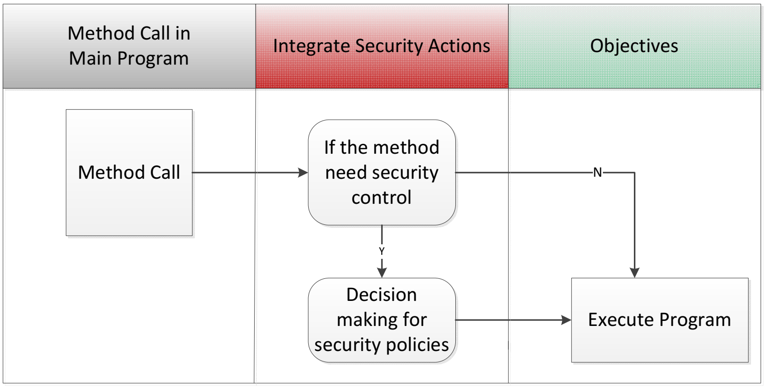

Method Call in Main Program
Integrate Security Actions
Objectives
Method Call
If the method need security control
N
Y
Decision making for security policies
Execute Program

Fig. 2. Common dynamic enforcement

Currently, AOP has support multiple language, such as Java, C, C++, most importantly, some researchers have implemented AOP for sensors [11], that provides us the feasible enforcement tools. AOP has dynamic injection mechanism which is mainly replace the original method by dynamic agent rewriting father class and intercepting messages to hacking program.

In AOP, there are some notations need to be explained. [Aspect], is represented the special objects modeled by cross-cutting concerns, which is different from the direct concerns of the other network program logics; [Pointcut], is to identify where the Aspect is be executed; [Advise], the action details in Aspect ; [Jointpoint], is the execution points which is corresponding to pointcut.

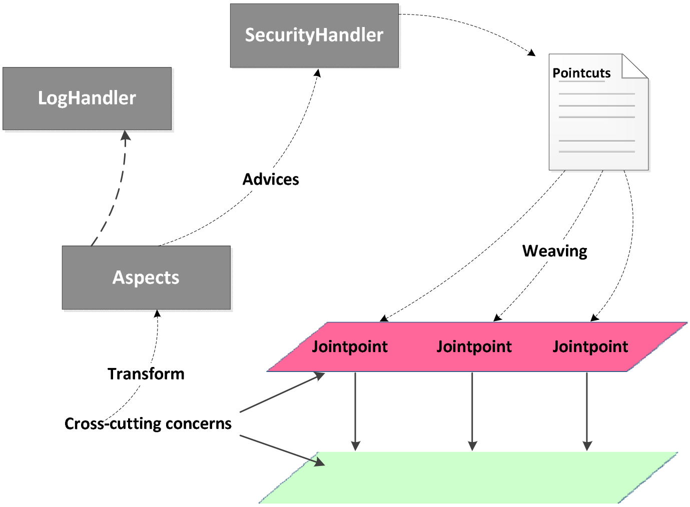


Fig. 3. Flow of aspect programming

The process of the integration has typical three steps, as shown in figure 3:

(1) Aspects decomposition: analyze the security requirement to propose the aspect concern points(as called pointcut). That means locating the interception points of target program which are waiting for security actions being intercepted.
(2) Concerns realization: realize the concerns (as called advises). The concerns mean the details of aspects in which the security actions are included.
(3) Aspects integration: dynamic inject aspects into network programs at the pointcut location. Thus the original programs are transferred into aspects involved programs. After this step, the objective program is generated. The security policies are embedded inside the network programs.

In it, we need to set up the corresponding interceptor which is a major way to interrupt the program and dynamic weave the policies into program. In our scheme, the interceptor mainly includes three types, Before, After, Around, which represent when to execute the security actions on the pointcuts. “Before” means the security measures should be enforced before the pointcut execution, “After” means the opposite way, “Around” means enforce the action before or after the pointcut execution.

3.4 Security Policies Categorization

In order to help judgment, we need to classify the security measures firstly. In this paper, we mainly classify security measures into three levels, level 1, level 2 and level 3. The level is higher, the security measures are stronger.

Table 1. Security Measures classification

Security Level	Effectiveness	Security Measures(some examples)
Level 1	Basic security protection	1. Optional encryption (or simple encryption algorithms such as RC5)
		2. Basic authentication (simple authentication protocols such as HMAC)
		3. No Integrity
Level 2	Medium security protection	1.Common encryption (such as TEA- Tiny Encryption Algorithm)
		2.Common authentication protocol(RSA 1024)
		3. Integrity checking (ECDSA 160)
Level 3	Strong security protection	1. Strong encryption(such as RC6)
		2. Strong authentication protocols(RSA 2048)
		3. Strong integrity checking(ECDSA 224)

4 Simulation and Analysis

- **Adaptation Simulation**

To evaluate the efficiency of MDPAS, we simulated it for 100 nodes generated in MATLAB. The parameters are set as shown in table 2. The deployment of nodes is random as shown in figure 4. Among the nodes, there are some compromised nodes which are supposed to launch the Sybil attacks[12] to the network.

Table 2. Paremeters

Configuration	Value
Total Area	100m × 100m
Number of nodes	100
Initial Energy	0.5J
Data rate	300 kbps
Transmission Range	10m
Packet size	48 bytes

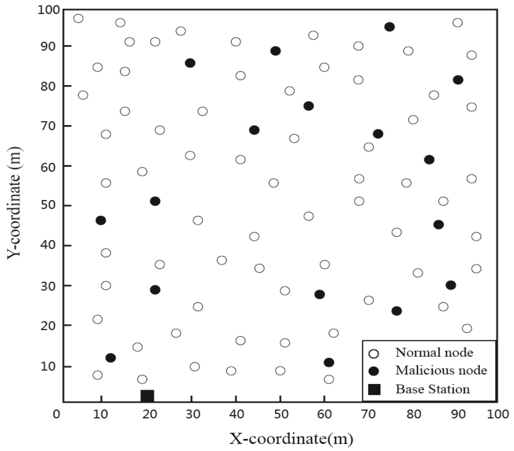


Fig. 4. Radom deployment of sensor nodes in an area of 100m×100m

We set different scenarios by the Sybil attacks, different percentage of compromised nodes means the security environment is different. In our simulation, we run 1200 rounds of hop to hop routing processes. As shown in figure 5, routing with the adaptive security can have much more alive nodes lasting in the networks than without adaptive security.

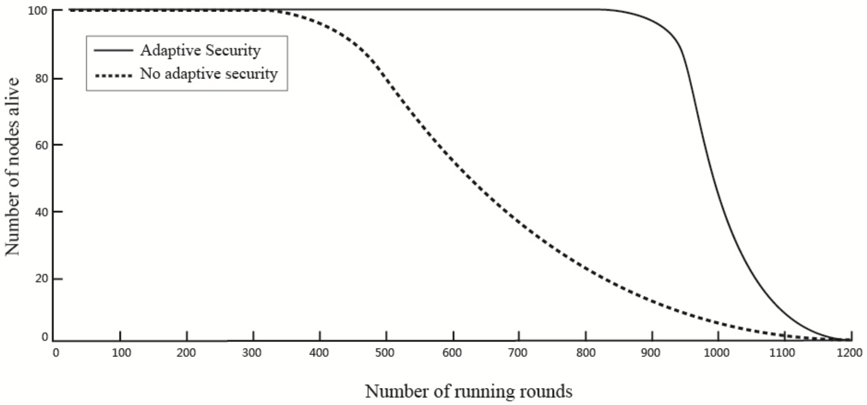


Fig. 5. Comparison of alive nodes

In order to test adaptive security ability, we set different security scenarios. We run the test again for 1200 times. From 0 to 400 rounds, we set zero percentage of compromised nodes, at round 400, compromised nodes rate is set to be 20%, at round 800, the compromised nodes rate is set to be 50%. According to our adaptive scheme MDPAS, it can dynamically adjust the security policies to meet the current environment requirements.

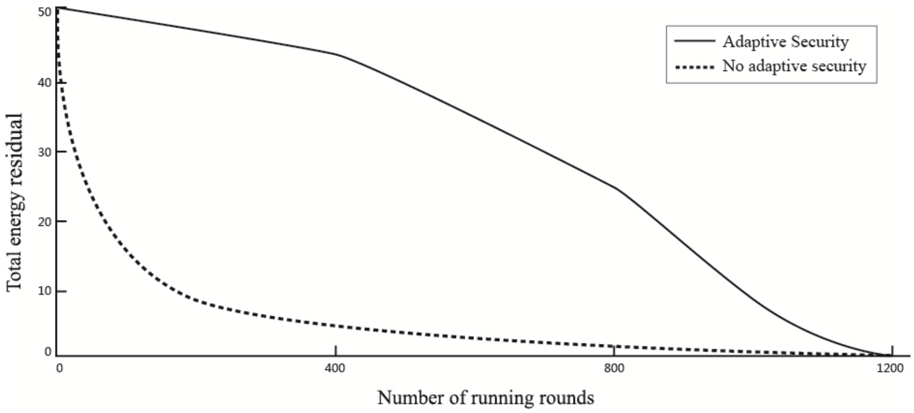


Fig. 6. Comparison of total energy residual

From the simulation result, as shown in figure 6, we can observe that from 0 to 400 rounds, the adaptive security loss less energy than non adaptive security since non adaptive security mechanism adopt the highest security policies from initial time to the end, while in the initial time, adaptive security mechanism just use lightest security measures; after 400 rounds, since the security scenario changes, the policies will be also changed to fit the current situation, then the security level is enhanced and energy consumption increases, so as after 800 rounds.

From the simulation result, we can conclude that the proposed adaptive security framework is effective.

Acknowledgments. This research was supported in part by National Natural Science Foundation of China (No.61100192), Research Fund for the Doctoral Program of Higher Education of China (No.20112302120074), grant No.10451805707004183 from Guangdong Province Natural Science Foundation and Shenzhen Strategic Emerging Industries Program under Grants No. JCYJ20120613151032592 and No. ZDSY20120613125016389. The authors thank the reviewers for their comments.

5 Conclusion

In this paper we propose an adaptive security framework for sensors in IoT, which provides dynamic confidentiality, authenticity and integrity in the networks with relative suitable overhead by context aware computing, decision making and dynamic enforcement of policies. We employ Markov Decision Process to make the decisions of security actions and adopt aspect-oriented programming technique to enforce the security policies dynamically in the working networks. We made simulation and the performance is encouraging.

References

1. Baldauf, M., Dustdar, S., Rosenberg, F.: A survey on context-aware systems. International Journal of Ad Hoc and Ubiquitous Computing 2(4), 263–277 (2007)
2. Santos, A.C., et al.: Challenges in the Development of Context-Inference Systems for Mobile Applications. International Workshop on Programming Methods for Mobile and Pervasive Systems (PMMPS), colocated with Pervasive (2010)
3. Min, J.-K., Cho, S.-B.: A hybrid context-aware wearable system with evolutionary optimization and selective inference of dynamic bayesian networks. In: Corchado, E., Kurzyński, M., Woźniak, M. (eds.) HAIS 2011, Part I. LNCS (LNAI), vol. 6678, pp. 444–451. Springer, Heidelberg (2011)
4. Mehra, P.: Context-aware computing: beyond search and location-based services. IEEE Internet Computing 16(2), 12–16 (2012)
5. Perera, C., Zaslavsky, A., Christen, P., et al.: Context aware computing for the internet of things: A survey (2013)
6. Stere, P., Cuppens, F., Cuppens-Boulahia, N., Garcia-Alfaro, J., Toutain, L.: Dynamic deployment of context-aware access control policies for constrained security devices. Journal of Systems and Software 84(7), 1144–1159 (2011)
7. Piotr, M., Magill, S., Hicks, M., Srivatsa, M.: Dynamic enforcement of knowledge-based security policies. In: 2011 IEEE 24th Computer Security Foundations Symposium (CSF), pp. 114–128. IEEE (2011)
8. Aloulou, H., Loulou, M., Kallel, S., Hadj Kacem, A.: RDyMASS: Reliable and dynamic enforcement of security policies for mobile agent systems. In: Garcia-Alfaro, J., Navarro-Arribas, G., Cuppens-Boulahia, N., Roudier, Y. (eds.) DPM 2009 and SETOP 2009. LNCS, vol. 5939, pp. 237–252. Springer, Heidelberg (2010)
9. Rafailidis, F., Panagos, I., Katsaros, P., et al.: Inlined monitors for security policy enforcement in web applications. In: Proceedings of the 17th Panhellenic Conference on Informatics, pp. 75–82. ACM (2013)
10. Singh, S.N., Singh, M.P.: Aspect Oriented And Object Oriented Software Using Java Programming Tools. Golden Research Thoughts 2(2) (2012)
11. Siegmund, N., Rosenmuller, M., Moritz, G., et al.: Towards robust data storage in wireless sensor networks. IETE Technical Review 26(5), 335 (2009)
12. Kumar, R.N., Bapuji, V., Govardhan, A., et al.: An Improvement to Trust Based Cross-Layer Security Protocol against Sybil Attacks (DAS). Computer Engineering and Intelligent Systems 3(7), 62–70 (2012)

Accurate Recommendation Based on Opinion Mining

Xiu Li, Huimin Wang, and Xinwei Yan*

Tsinghua University, Department of Automation, Lishuistr. 2279, Shenzhen, China

Abstract. Current recommender systems are mainly based on customers' personal information and online behavior. We find that those systems lack efficiency and accuracy. At the same time, we observe the large amount of review data with exponential growth. Based on this observation, we propose a recommender system based on opinion mining. With text mining method we extract the opinion related information from the massive reviews. We analyse the linguistic information and design a two-layer selection algorithm to find the most suitable products for customers. The experiment shows our method has great accuracy, fleasibility, and reliablity.

Keywords: e-business, opinion mining, recommender system.

1 Introduction

The Internet Era brings convenient information service, promotes the rapid development of E-commerce, and also has a profound impact on people's way of life. Both information acquisition and shopping consumption have generally turned to the online. China has the largest online market in the world, but consumers are always limited by cognitive ability and the immature information search behavior confronted with the complicated online market. As one of the most appropriate ways, accurate recommendation can help customers to quickly find the products they need. Opinion mining has great potential to be used in personal recommender system to significantly improve: i) accuracy by analyzing individual review data, and ii) reliability by considering numerous opinions of customers. However, designing a recommender system based on opinion mining is quite difficult, which should:

- Correctly recognize the sentiment expression in complicated Chinese text, extract feature words, adjectives and adverbs, and match them up.
- Accurately calculate quantified sentiment strength of semantic features, and output a normalized numerical value in certain dimensions.
- Appropriately recommend products according to customers' review information.
- Obtain good generalization ability, high computational efficiency.

To the best of our knowledge, no prior system can satisfy all these goals simultaneously.

* Corresponding author.

H. Sun et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 329, DOI: 10.1007/978-3-319-12286-1_41

In this paper, we propose a new recommender system based on opinion mining. Firstly we establish a customer comments feature library and extract the semantic features by exploring the linguistic information. Next, we put up with a new method on fine-grained sentiment analysis to summarize opinion mining and statistics. Finally, indicators are generated and products to be recommended are selected by two-layer selection. The experimental results show that our system efficiently improves the accuracy and practicability of recommendation.

2 Related Works

The current recommender system of E-commerce shopping websites in China mainly relies on the purchase history, which can be divided into three categories [1]:

1* Collaborative filtering [2]. It explores the adjacent customers first and recommends what they love to the target customer. It dose not consider the product attributes which are pivotal in the system.

2* Content-based Recommendation [3,4]. It extracts the product features and generates feature vectors. Then it views customer shopping history and recommends similar product they had bought based on distances between feature vectors. It dose not consider customers' attitude towards the product, which limits the accuracy of recommendation.

3* Knowledge-based Recommendation [5]. It requires customers propose the demand first and the whole process is strongly interactive. Obviously, it can't obtain new customers initiatively and it's time consuming.

To solve the problems, more and more scholars began to research in E-commerce recommender system from various perspectives. Some of them are from the perspective of the users' opinion mining. Such a system based on opinion mining can be divided into three parts: Sentiment expression recognition, sentiment analysis and recommendation algorithm.

The process of sentiment expression recognition is based on Chinese Natural Language Processing (NLP) tools which can achieve word segmentation, POS tagging, named entity recognition and anaphora resolution. To extract sentiment expressions, key words related a certain theme have to be identified. Although there are many methods that focus on Chinese key words extraction, almost none of them consider the character of topic relativity. Jianlin Zhang and Qianli Shen [6] apply Dunning's [7] possibility algorithm to identified credit related key words, but the process is too complicated and time-consuming. Wenhua Wang and Yanhui Zhu try to extract the product attribute words and opinion words with machine learning method.

Sentiment analysis aims at recognizing words, sentences or document's sentimental polarity. Benefiting from the development and maturity of the technology in natural language processing and machine learning, it becomes possible to widely employ sentiment analysis on Chinese texts. Existing studies on sentiment analysis are mainly focusing on the task of determining word-level and sentence-level. Yanlan

Zhu and Jing Min use the semantic similarity between two words in HowNet library to distinguish sentence polarity [8]. Zhenyu Wang and Zeheng Wu combine point mutual information with semantic similarity to improve performance [9]. Hongwei Wang and Lijuan Zheng apply machine learning to get the sentimental contribution degree of sentence [10]. These methods rely on the accuracy of library, but there is no library directed to E-commerce.

The studies on recommendation algorithm are mainly focus on the three methods mentioned above.

3 Methodology

Our system mainly covers two parts—reviews data reproducing and recommendation algorithm. With the use of Fudannlp, a famous open source tool in the field of Chinese language processing, we apply SVM to extract attribute words and their opinion words (Sec.3.1). Then we calculate the sentimental strength of sentiment expressions (Sec.3.2). After that, we select products through two layers (Sec.3.3). More specific process is shown in the flow chart as follow:

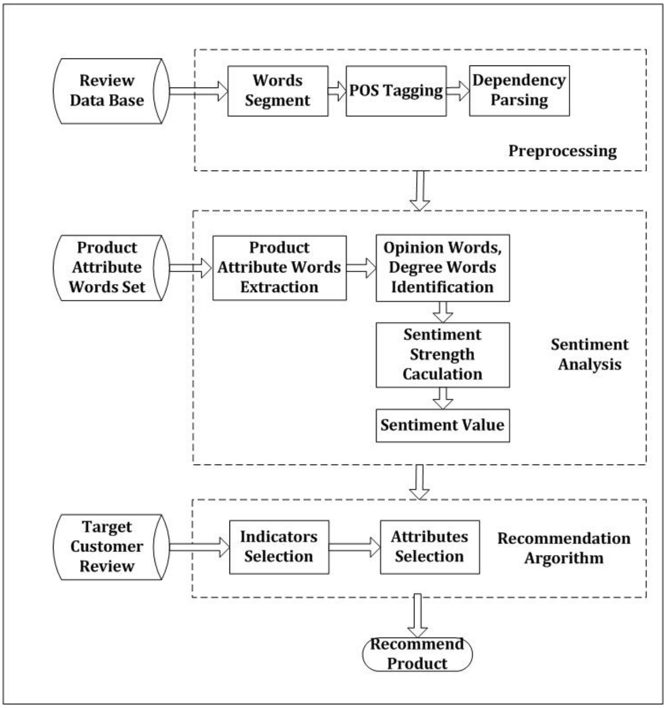


Fig. 1. The flowchart of the recommender system

3.1 Key Words Extraction

For every product attribute, the key words include attribute word, related opinion words, negative adverbs, and degree adverbs. The attribute word should be one of the features of the product. Recently, the accuracy and recall of attribute words extraction are relatively low [6]. To improve the performance, we refine the process and divide it into several steps:

Candidate Words Selection: Although Chinese sentences are complex and diverse, they show some statistic characteristics when limited to product reviews. We randomly crawled 4,000 customer reviews from Jingdong which is one of the chief online shopping malls in China. After text data preprocesses including Chinese word segment, POS tagging, and dependency parsing, we count 90% of attribute words are noun and 75% of them act as the subjects of the sentences. This is highly in accordance with the Chinese daily way of speaking. Therefore, we mainly extract the words with the POS tagging "noun" and parsing feature "subject" as candidate attribute words.

Product Attribute Words Extraction: Based on candidate attribute words, we go further more to choose to credit indicator related words as attribute words by matching the product attribute words set. More details, we establish word sets W_1... W_n, and each of them contains the common used words of one aspect of product. To prune the candidate attribute words, we calculate every candidate's semantic distance with every element w in W_1... W_n using the method based on HowNet and normalize it as Equation (1), whose values are between 0 and 1.

$$Sem(w,s)=\frac{d(-\frac{1}{2}S(w,s)+1)}{S(w,s)+d} \tag{1}$$

Where $S\ (w,\ s)$ is the space distance of Space Vector Model between w and s, d is a controllable parameter, and s is a candidate attribute word. $if\,\exists i\in\{1,2,L,n\},\ s.t.\ Sem(s,Wi)\geq t$, where t is a threshold, then s is a chosen attribute word.

Opinion Words, Degree Adverbs, and Negative Adverbs Identifying: for every product attribute word, we identified its opinion words, negative adverbs, and degree adverbs with SVM. With *one versus rest*, one of the most widely used methods of SVM multi-classifier, we classify the words excluding the attribute words in one piece of review to opinion words, negative adverbs, degree adverbs and other words. For the convenience of description, we noted one attribute word's opinion words, degree adverbs, and negative adverbs as its related words. Furthermore, to facilitate classification, we quantify the relationship between one attribute word and its related words. Under the comprehensive analysis of the grammar features, the orders and the distances between attribute words and their related words, we make **feature selection rules** as follow:

1* Part of speech of the related words. According to the 39 kinds of outputs of POS Tagging with FudanNLP-1.6.1, we number the different outputs from 1 to 39.

2* Dependency parsing's result of the related words. There are 22 kinds of syntactic relationships of Dependency Parsing in FudanNLP-1.6.1, we number them from 1 to 22. E.g. we label the *Subject* to 1.

3* the distance between an attribute word and its related word. We calculate the number of characters excluding spaces between the attribute words and their related words.

4* the order of an attribute word and its related word. If the attribute word is preceded with its related word, we label this dimension of feature as 1, else 0.

5* is there a punctuation. We will label the feature as 0, if there is no punctuation between the attribute word and its related word. Else, if there is a period, label 1; if there is a comma, label 2; if there is a apace, label 3; else, label 4.

6* is there another attribute word. If there is another attribute word between the attribute word and its related word, we label this feature as 1, else as 0.

7* is there an opinion word. If the attribute word has opinion words, the dimension of feature will be labeled as 1; else, labeled as 0.

8* is there a degree adverb. If the attribute word has degree adverbs, the dimension of feature will be labeled as 1; else, labeled as 0.

9* is there a negative adverb. If the attribute word has negative adverbs, the dimension of feature will be labeled as 1; else, labeled as 0.

Training Classifiers. We choose the reviews from www.jd.com as training set after manually annotating. Specifically, we manually extract the product attribute words and identified their related words firstly, after that we will get the set of evaluation unit $G = \{(k1,d1,p1,o1), \ldots,(kn,dn,pn,on)\}$ for every piece of review, where n is the number of attribute words in the review, *ki, di, pi, oi* represent attribute words, degree adverb, negative adverb, and opinion word.. Then we quantify the relation features between attribute the attribute word and its related words. For example, there is an evaluation unit *(k,d,p,o)=(*布料, 非常, 不, 好*)* extracted from the review—这件衣服的布料非常不好, which means *the fabric of this dress is very bad*. The manual annotation is as follow:

Table 1. Manual annotation

Match or not	Attribute word	Related words	feature
1	*k*	*d*	2 4 0 1 0 0 0 0 0
1		*p*	3 5 4 1 0 0 0 1 0
1		*o*	3 6 6 1 0 0 0 1 1

3.2 Sentiment Analysis

We analyze every opinion word's semantic orientation with the method of word's similarity computing based on Chinese Thesaurus -*Tongyici Cilin*, which not only contains one word's synonyms, but also a certain number of its similar words. This dictionary is compiled in 1983, and has not been updated from then on. At the same time, the *Tongyici Cilin extension* keeps the original edition's three layer classification system, and adds two layers to be further sub-classes. With the final five layers classification, the words in the dictionary shows the good hierarchical relationships, which can be expressed as shown below, L1 means the first layer and the remains are in the same way.

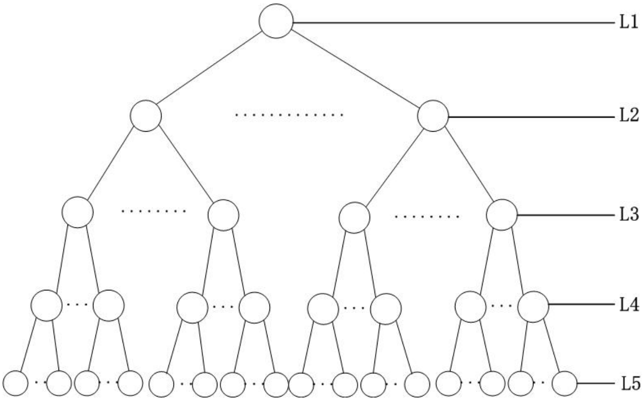


Fig. 2. Hierarchical relationships in *Tongyici Cilin extension*

Now the first layer will be C, the second will be b, the third will be 02 the fourth will be A, and the fifth will be 01. We definite *S (w1, w2)* as the similarity between *w1* and *w2* and its value belong to 0 to 1. The bigger the value is, the higher the similarity is. Let n represents the number of the layers, and the k is the distance of the two branches. The value of *S (w1, w2)* can be discussed in the following circumstances:

If *w1* and *w2* are not in the same search tree:

$$S\,(w1,\,w2) = f$$

If *w1* and *w2* are both in the 2nd layer of a same search tree:

$$S(w1,w2) = a\frac{n\text{-}k+1}{n}\cos(\frac{n\pi}{180})$$

If *w1* and *w2* are both in the 3rd layer of a same search tree:

$$S(w1,w2) = b\frac{n\text{-}k+1}{n}\cos(\frac{n\pi}{180})$$

If *w1* and *w2* are both in the 4th layer of a same search tree:

$$S(w1,w2) = c\frac{n\text{-}k+1}{n}\cos(\frac{n\pi}{180})$$

If *w1* and *w2* are both in the 5th layer of a same search tree:

$$S(w1,w2) = d\frac{n\text{-}k+1}{n}\cos(\frac{n\pi}{180})$$

Parameter *a, b, c, d, e, f* will be determined from the experimental result. One word may have several code, we choose the biggest value of the similarity. For example, the codes of word “骄傲（pride）” are “Da13A01” and “Ee34D01”. And the codes of the word “仔细（careful）” are “Ee26A01” and “Ee28A01”. There will 4 values of the similarity between the two words, there are 0.1, 0.1, 0.483920, and 0.51007, but we final choose the biggest one 0.51007. What is more, we get the similarity between word “美丽（beautiful）” and other words in Tab.2:

Table 2. The similarity between word “美丽（beautiful）” and other words

Word	Similarity
漂亮(beautiful)	1.000000
丑陋(ugly)	0.161666
可爱(cute)	0.582177
灿烂(splendid)	0.478922

For one evaluation unit *(k, d, p, o),* we get the sentiment of the opinion word *o* according to Formula (2) as follow:

$$Ori(o) = w_n(p) \cdot w_{\mathrm{d}}(d) \cdot (\frac{\sum_{i=1}^{n} S(wp_i, o)}{n} - \frac{\sum_{j=1}^{m} S(wn_j, o)}{m}) \qquad (2)$$

Where wp_i and wn_j separately are basic words of positive and negative word set, *n* and *m* separately are the number of basic words of positive and negative word set. W_n is the weight of the negative adverb, and W_d is of the degree adverb. If there is no negative adverb *n*, W_n is equal to 1. Else W_n can be obtained by Formula (3):

$$W_n = \max_{i \le m}(S(wn_i, n)) \qquad (3)$$

Where wn_i is the word of negative adverb set, and *m* is the number of the words in the set.

If there is no degree adverb *n*, W_d is equal to 1. Else W_d can be obtained by Formula (4):

$$W_d = 1.1 \max_{i \le m}(S(wdi, d)) \qquad (4)$$

Where *wdi* is the word of highest degree adverb set, and n is the number of the number of the words in the set. In fact, degree word can be classified to three types — high, intermediate, and low. We mainly build the highest degree adverb set, with which degree adverb d is compared. .

Since the value of w_n, w_{d} and $S(w1, w2)$ are belong to [0, 1], the value of the sentimental orientation of opinion words should range from -1 to +1, and the bigger the value is, the stronger the sentimental orientation is and the higher the score of the indicator corresponding to the attribute word is.

3.3 Recommendation Algorithm

So far we have obtained the subjective feedback from customers. Next, we propose a two layers matching algorithm. We establish an indicator system as the upper layer with clustering method. Then in the sublayer, for every single review we select attributes with high customer scores as feature attributes. After that, we search for products through these two layers according to customers' opinion.

3.3.1 Attributes Clustering

Customer review scores can reflect the quality of products, but the amount of attribute words is huge and dimensionality reduction should be done. In order to simplify the search complexity, similar attributes should be combined into an indicator and the number of indicators should be appropriate. We use K-means clustering method to obtain indicators from attribute words. We try *1* to *25* as K value which represents the number of class centers. It shows that when K=*12*, the classes have best performance on the within class scatter and the between class scatter.

So we select one word that summarizes the class as the indicator and set the indicator system with *12* indicators which are shown below.

Indicators: *Quality, Performance, Function, Price, Advertising authenticity, Online service, After-sale, Logistics service, Transaction security, Transaction convenience, Transaction frequency, Transaction value.*

Every review can be quantized as a vector with 12 dimensions: $R=\left(r_1, r_2 \cdots r_{12}\right)$. And every product can be quantized as: $\overline{R}=\left(\overline{r_1}, \overline{r_2} \cdots \overline{r_{12}}\right)$, where $\overline{R}$ is the mean of all reviews of the product.

3.3.2 Attributes Selection

As a product to be recommended, we select its advantageous attributes as feature attributes and use them to represent the product. Feature attributes should be satisfied with following conditions:

1* High customer score. The scores of feature attributes should be higher than a threshold σ_s which can be set manually.

2* Enough review rate. There should be enough customers in favour of it when setting a feature attributes. Attributes with review rate under σ_r are not in the area of concern.

For every product, we establish a set of attributes $A=\left\{a_1, a_2 \cdots a_n\right\}$ which can be summarized as an indicator set $I=\left\{i_1, i_2 \cdots i_n\right\}(\mathrm{n} \leq 12)$.

3.3.3 Product Matching

We assume that customers express what they care about by comments. When a customer writes a review which contains several attributes $A'=\left\{a_1', a_2' \cdots a_n'\right\}$ in certain indicators $I' = \{i'_1, i'_2 \cdots i'_n\} (\mathrm{n} \leq 12)$. We first search for the same kinds of products with the same indicator set, which means $I = I'$. After that we search for products with $A' \subseteq A$ in the candidate products. The more details the customer review contains, the more accurate the recommendation will be.

3.4 Experimental Analysis

Text Mining Performance Test. In the text mining, if the related word is matched, we label 1 as matching information, if not, label 0. Then we get rid of the text information and set the matching information and quantified feature characters as input to train the classifiers. We randomly choose 70% of the manually annotated data, around 700 pieces of reviews, as training set, the remains as testing set. To evaluate the performance of the classifier, we use the test data and calculate the accuracy, recall, and F1-means, and the result is shown in Tab. 3.

Table 3. Evaluation-unit extracting result

Precision	Recall	F1-means
85.05%	80.47%	82.69%

It is obvious that text mining performance is improved significantly.

Recommendation Performance Test. Due to the particularity of the system mechanism, it is difficult to directly test its performance. We assume that when customers review with sale slips, they are satisfied with the product. We randomly select 100 reviews with slips and compared their products and products our system recommended, the result is shown in Tab.4.

Table 4. Evaluation-unit extracting result

Same brand	Other brands with close scores	Other brands with far different scores
34%	61%	5%

It shows that 95% of customers get effective recommendation.

4 Conclusions and Future Work

We present an efficient but surprisingly simple e-business recommender model based on opinion mining. Text mining performance is improved significantly and the refined data provides a good foundation of other related researches. We design every detail with actual mechanism and in the whole process of modeling, we maintain a good objectivity.

Limitations. The theme related key words' extracting depends on word sets, which affects the portability of the model. The recommendation algorithm may be combined with existing algorithms to improve the performance one step further.

Future Works. Deep learning combined by POS and parsing character can be used in theme related words extracting to improve its efficiency.

References

[1] Adomavicius, G., Tuzhilin, A.: Towards the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. IEEE Transactions on Knowledge and Data Engineering 17(6), 734–749 (2005)

[2] Mooney, R.J., Roy, L.: Content-based book recommending using learning for text categorization. In: Proceedings of the Fifth ACM Conference on Digital libraries, DL 2000, New York, pp. 195–204 (2000)

[3] Zhou, J., Luo, T.: Towards an Introduction to Collaborative Filtering. In: Computational Science and Engineering, pp. 576–581. IEEE, Los Alamtios (2009)

[4] Si, L., Jin, R.: Unified Filtering by Combining Collaborative Filtering and Contend-Based Filtering via Mixture Model and Exponential Model. In: Knowledge Manage, pp. 156–157. ACM, New York (2004)

[5] Jannach, D., Zanker, M., Felfering, A., et al.: Recommender Systems: An Introduction. Cambridge University Press (2011)

[6] Zhang, J.L., Shen, Q.L., Wu, J.Y.: E-Businessmen credibility mining based on web reviews. In: Proceedings of 2nd International Conference on E-Business and E-Government, pp. 5955–5961. ICEE, Shanghai (2011)

[7] Dunning, T.: Accurate methods for the statistics of surprise and coincidence. Computational Linguistics 19(1), 61–74 (1993)

[8] Zhu, Y.L., Min, J., Zhou, Y.Q., et al.: Semantic Orientation Computing Based on HowNet. Journal of Chinese Information Processing 20(1), 14–20 (2006)

[9] Wang, Z.Y., Wu, Z.H., Hu, F.T.: Words Sentiment Polarity Calculation Based on HowNet and PMI. Computer Engineering 38(15), 187–189 (2012)

[10] Wang, H.W., Zheng, L.J., Yin, P., et al.: Chinese Web Reviews Sentiment Polarity Classification Based on Sentence Level Sentiment. Journal of Management Sciences in China 16(9), 64–74 (2013)

Reliability Analysis on Parallel System with N Element Obeying Exponential Distribution

Lu-Xiong Xu[1,2], Chao-Fan Xie[3], and Lin Xu[1,4,*]

[1] The Institute of Innovative Information Industry,
Fuqing Branch of Fujian Normal University, China
[2] The School of Mathematics and Computer Science,
Fuqing Branch of Fujian Normal University, China
[3] Network Center, Fuqing Branch of Fujian Normal University, China
[4] The School of Ecomomic, Fujian Normal University, China
xulin@fjnu.edu.cn

Abstract. The modern computer and network system, is composed of tens of thousands of components. Any faults in components can affect reliability of the whole system. In the process of running, elements will gradually age and failure rate gradually increase. However, the rate at which these elements age is not synchronous. Thus, in which cases should the elements be upgraded or replaced and in which cases should the elements whose performance is still relatively good be retained, are hard job to handle. In this paper, the author attempts to make an extreme analysis on the reliability of parallel system with N element obeying exponential distribution to improve the system update.

Keywords: reliability, parallel system, deletion, exponential distribution, system update.

1 Introduction

1.1 Reliability Introduction

Reliability theory was a newly emerged interdisciplinary subject in 1960s and analyzes the probalility of random events which characterize the specified function of product. It is established on the basis of probability theory, which is an area of study focused on machine maintenance[1].With the development of reliability theory, it gradually needs much frontier knowledge and tools in mathematics, while reliability mathematics has laid a food foundation for it. In practical reliability problems, the mathematics used can be divided into two categories: probability model and statistical model. Probability model infers the reliability indices of system on the basis of system structure and life distribution of components; while statistical model evaluates and tests the life of components or system on the basis of observed data. In the paper, statistical model is applied. Currently, the main researches focus on the reliability

[*] Corresponding author.

H. Sun et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 329, DOI: 10.1007/978-3-319-12286-1_42

indices of system and optimal detecting time which is determined by reliability indices to avoid the occurrence of faults and reduce the losses caused by faults, such as literature[2],[3],[4]. In this paper, the author attempts to make an extreme analysis on the reliability of parallel system with N element obeying exponential distribution to improve the system update.

1.2 The Definition of the Main Indicators of Reliability

(1) Reliability

The definition of reliability *R(t)* [5]: it is the probability that product completes the required function under the specified conditions and within the prescribed time.

If the life distribution of product is $F(t)$, $t > 0$, the reliability $R(t) = P(T \geq t) = 1 - F(t)$ .This is a function of time(*t*), so it can be called as reliability function. To the components obeying exponential distribution λ, its reliability is $e^{-\lambda t}, t \geq 0$.

(2) Failure rate

Failure rate $\lambda(t)$: It is the probalility of occuring failure in the unit of time after product has worked a period of time(*t*). According to reliability theory, $\lambda(t) = \frac{f(t)}{1 - F(t)}$, when $t > 0$, the failure rate of exponential distribution is constant λ.

(3) System parameter specification

A : represents normal working events of system.

A_i : represents normal working events of the element i.

λ_i : represents failure rate of the element i.

$R_s(t)$: represents system reliability,that is, $P(A) = R_s$.

$R_i(t)$: represents reliability of the element *i*, that is, $P(A_i) = R_i$.

Parallel system: It is a system consisting of n components. As long as one of these elements works, the system can work.; only when all the units fail, the system would fail. The logical block diagram of parallel system with n element as follows(figure 1).

According to the property of probability, the normal working probability of system $P(A) = P(\bigcup_{i=1}^{n} A_i)$ is as follows:

$$R_s(t) = 1 - \prod_{i=1}^{n}(1 - R_i(t)) = 1 - \prod_{i=1}^{n}(1 - e^{-\lambda_i t}) \tag{1}$$

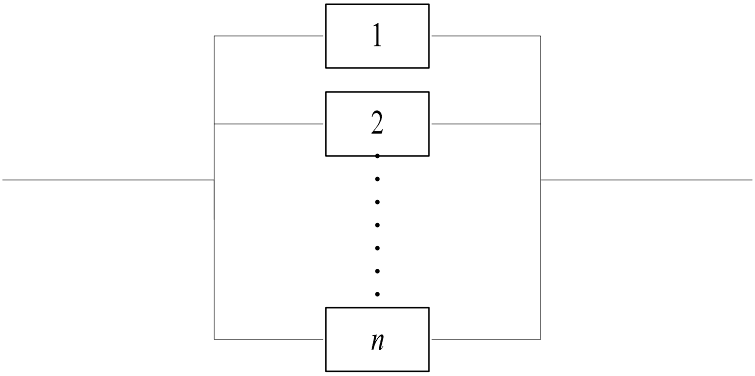


Fig. 1. The logical block diagram of parallel system

2 Model Analysis

2.1 Peformance Constraint

According to the definition of parallel system reliability, when the overall reliability of N elements is fixed, the model is as follows:

$$\begin{aligned} &\min R_s(t) = 1 - \prod_{i=1}^{n} (1 - e^{-\lambda_i t}) \\ &s.t. \sum_{i=1}^{n} \lambda_i = c, \lambda_i > 0 \end{aligned} \tag{2}$$

C is the sum of failure rate.

This model is equivalent to the following one:

$$\begin{aligned} &\max \quad (\prod_{i=1}^{n} (1 - e^{-\lambda_i t})) \\ &s.t. \sum_{i=1}^{n} \lambda_i = c, \lambda_i > 0 \end{aligned} \tag{3}$$

If multivariable differential calculus and lagrangian multiplier are adopted here, the process of caculating conditional extremum will be complex. Thus, the author attempts to adopt Cauchy inequation here and conclude as follows:

$$\prod_{i=1}^{n} (1 - e^{-\lambda_i t}) \le \left(\frac{\sum_{i=1}^{n} (1 - e^{-\lambda_i t})}{n} \right)^n = \left(1 - \frac{\sum_{i=1}^{n} e^{-\lambda_i t}}{n} \right)^n \tag{4}$$

Using the Cauchy inequation once again, the following can be concluded:

$$\left(1-\frac{\sum_{i=1}^{n} e^{-\lambda_i t}}{n}\right)^{n} \leq \left(1-\sqrt[n]{\prod_{i=1}^{n} e^{-\lambda_i t}}\right)^{n} = \left(1-e^{-\frac{ct}{n}}\right)^{n} \quad (5)$$

The two inequalities above can be turned into equal, but this is only true when $\lambda_1 = \lambda_2 \cdots = \lambda_n = \frac{c}{n}$. Thus, when failure rate of all components are equal, the reliability of system would reach minimum value. In other words, when the sum of failure rate is certain, the reliability of selecting products with the same quality cannot be better than the reliability of choosing a relatively poor one from a good parallel product.

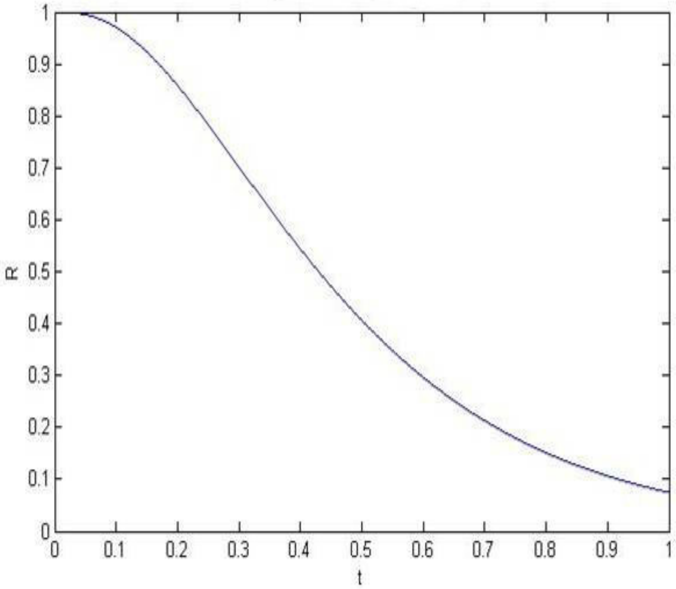


Fig. 2. the envelope curve of the lowest reliability In practice, in the case that n elements in parallel are all from the parent which is obeying parameter λ , the abrasion and depreciation of components is growing over time, so the failure rate is in an increasing trend, thus c is increasing. Now assuming *c* is the linear function of *t*, the envelope curve of the lowest reliability in accordance with the time can be concluded as follows(figure 2):

In practice, in the case that n elements in parallel are all from the parent which is obeying parameter λ , the abrasion and depreciation of components is growing over time, so the failure rate is in an increasing trend, thus c is increasing. Now assuming *c* is the linear function of *t*, the envelope curve of the lowest reliability in accordance with the time can be concluded as follows (figure 2):

$$R_{\min}(t) = 1 - (1 - e^{-\frac{at^2+bt}{n}})^n, a < 0, b > 0 \tag{6}$$

Wherein *a* represents wear rate, *b* is saturation mode when it is never used.

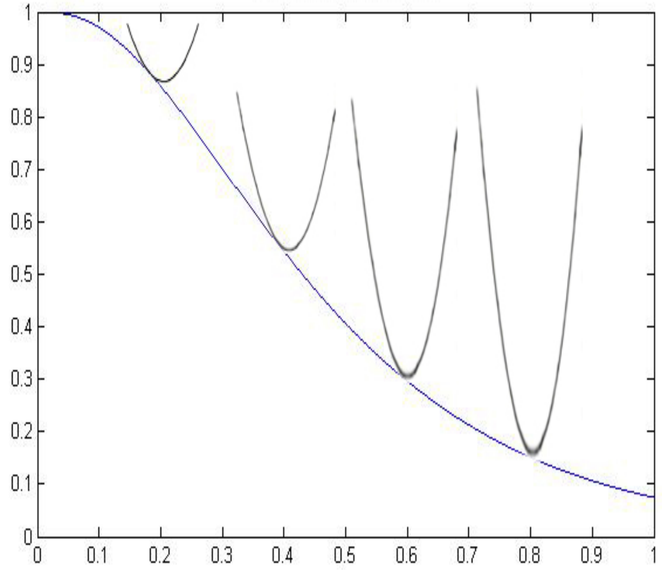


Fig. 3. System reliability and envelope curve

In the process of running, the curves of system reliability are always above the envelope curve, just as what imply in the following chart(figure 3). When the system components are operating and the failure rate reaches the minimum acceptable limits, according to the actual situation, if the curve of failure rate is above the envelope curve, then all the components of the system have to be replaced; if the curve of failure rate is still relatively distant from the envelope curve, then the components with the highest failure rate are needed to replace according to the economy principle. In reality,the system reliability and envelope curve can be getted by collecting data which is used for statistic to approching.

2.2 Relations between Cost and Failture Rate

Now the paper introduces the cost constraint,first we define the function of failture rate and corresponding costs, ,denoted as $\lambda = g(\mu)$,thereinto μ is expense, λ

is still failture rate. Here we learn the definition margin of microeconomics, define the concept of marginal reliability[6]. Marginal reliability: the mount of decreasing failure rate arouse by increasing one unit of cost.

In actual, designers can choose relatively low failure rate product when he add budget, so $g^{'}(u) < 0$. But with the increasing of cost, the amount of reducing failure rate is decreasing, thus there is $g^{''}(u) > 0$. Next we denote the inverse function of $\lambda = g(\mu)$ as $u = f(\lambda)$, since $\lambda = g(\mu)$ and $u = f(\lambda)$ are mutually inverse function, so $g^{'}(u) = \frac{1}{f^{'}(\lambda)}$. According to the characteristics of the marginal failure rate, we can get $(\frac{1}{f^{'}(\lambda)})^{'} < 0$, that is $f^{''}(\lambda) > 0$.Lastly, we define failure rate flexibility as the changing ratio of the failure rate divide changing ratio of cost, denote as E_λ.Its mathematical expression as following:

$$E_\lambda = \frac{\frac{du}{u}}{\frac{d\lambda}{\lambda}} = \frac{\lambda}{u}\frac{du}{d\lambda} \tag{7}$$

Failure rate flexibility can be divided into the following three types:

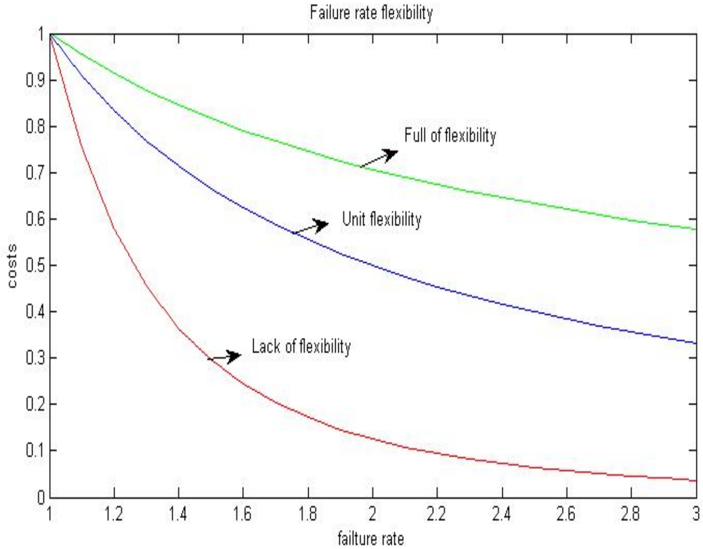


Fig. 4. Failure rate flexibility

(1) Lack of flexibility: $E_{\lambda} > 1$, for sophisticated originals,it requires a lot of R & D funding, therefore, the changing of failure rate ratio will cause large changing in the ratio of cost.

(2) Full of flexibility: $E_{\lambda} < 1$,for those who focus more on appearance innovation of product, its failure rate is full of flexibility because appearance innovaion need less cost.

(3) Unit flexibility: $E_{\lambda}=1$, for mature products and under full competition of market ,the majority of products are such cases.

the relation of three types of Failure rate flexibility shown in frigure(4).

2.3 Costs and Constraint

By definition of parallel system reliability,under the peformance and costs constraints, the model of n Element Obeying Exponential Distribution as below:

$$\min R_s(t) = 1 - \prod_{i=1}^{n} (1 - e^{-\lambda_i t})$$
$$s.t. \begin{cases} \sum\limits_{i=1}^{n} \lambda_i \leq c \\ \sum\limits_{i=1}^{n} f(\lambda_i) \leq u \\ \lambda_i > 0 \end{cases} \tag{7}$$

c represents the peformance constraint of the system, and u represents economic costs constraints of the system. Depending on the feature of $f(\lambda)$, firstly we consider the case of the failure rate flexible which is unit costs,based on formula (7): $E_{\lambda} = \frac{\frac{du}{u}}{\frac{d\lambda}{\lambda}} = \frac{\lambda}{u}\frac{du}{d\lambda} = 1$, learn by microeconomics,it represents the function of $u = \frac{1}{\lambda}$.The model is replaced by the following:

$$\max R(t) = \prod_{i=1}^{n} (1 - e^{-\lambda_i t})$$
$$s.t.\begin{cases} \sum_{i=1}^{n} \lambda_i \le c \\ \sum_{i=1}^{n} \frac{1}{\lambda_i} \le u \\ \lambda_i > 0 \end{cases} \tag{8}$$

Firstly we analyse feasible solution space, according to the Cauchy-Schwarz inequality:

$$(\sum_{i=1}^{n}\lambda_i)*(\sum_{i=1}^{n}\frac{1}{\lambda_i})\ge(\sum_{i=1}^{n}(\sqrt{\lambda_i}*\frac{1}{\sqrt{\lambda_i}}))^2 = n^2 \tag{9}$$

Thus we have:

$$u \ge \frac{n^2}{c} \tag{10}$$

$$u \ge \frac{n^2}{c} \Rightarrow \frac{n}{u} \le \frac{c}{n} \tag{11}$$

The inequalities above can be turned into equal, but this is only true when $\lambda_1 = \lambda_2 \cdots = \lambda_n = \frac{c}{n}$, formula(11) is a prerequisite for the existence of a feasible solution, also can be seen from the formula, in order to ensure the system have at least lowest peformance, the least costs is $u = \frac{n^2}{c}$,if $u < \frac{n^2}{c}$,the costs is insufficient to support the current system of minimum peformance. Having a feasible region of *u-c* relationship as shown in figure 5: and known by Cauchy inequality, when $\lambda_1 = \lambda_2 \cdots = \lambda_n$, system reliability reaches a minimum, so whether is this solution feasible?When $\lambda_1 = \lambda_2 \cdots = \lambda_n = \lambda$,the minimal reliability of the system is the function of λ and t:

$$R_{\min}(\lambda) = 1\text{-}\left(1 - e^{-\lambda t}\right)^n \tag{12}$$

the minimal reliability of the system is decreasing function of λ ,and known by the inequality constraints:

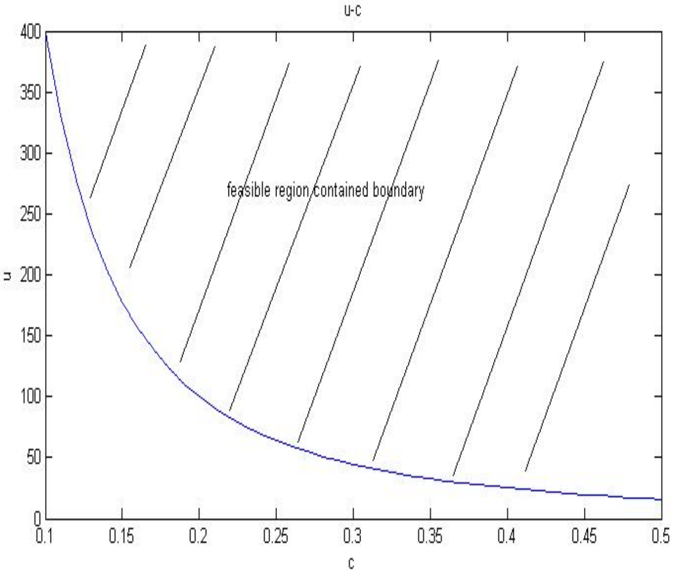


Fig. 5. Feasible solution space

$$\frac{n}{\lambda} \le u, n\lambda \le c \Rightarrow \frac{n}{u} \le \lambda \le \frac{c}{n} \tag{13}$$

λ can get to the maximum of $\frac{c}{n}$, therefore (3) can be reduced to the formula:

$$R_{\min} = 1\text{-}\left(1 - e^{-\frac{c}{n}t}\right)^n \tag{14}$$

The result is the same as only have peformance constraint.

3 Conclussion

On the condition that the sum of failure rate of system components is fixed, this paper attempts to make an extreme analysis on the reliability of parallel system with N element according to the cauchy inequality, and put foward an meaningful conclusion and related theory as a guidance. When the failure rate of all components are equal, the system reliability can reach the minimum value,also under the peformance and costs constraints,the result is same as only peformance constranit,but it have least costs to support its feasible solution space which can't become empty set. Therefore, in practice, it is better to choose a relatively poor one from a good parallel product than select products with the same quality, because the reliability of the former is better than that of the latter. In the process of running, according to the actual situation of system, if the curve of failure rate is above the envelope curve, then all the components of the system have to be replaced; if the curve of failure rate is still relatively distant from the envelope curve, then the components with the highest failure rate are needed to replace according to the economy principle.

Acknowledgement. This research was partially supported by the the School of Mathematics and Computer Science of Fujian Normal University, by the the Institute of Innovative Information Industry of Fujian Normal University,by the The School of Ecomomic of Fujian Normal University.

References

1. Luss, H.: An Inspection Policy Model for Production Facilities. J. Management Science 29, 101–109 (1983)
2. Zequeira Romulo, I., Berenguer, C.: On the Inspection Policy of a Two-Component Parallel System with Failture Interation. J. Reliability Engineering and System Safety 88(1), 99–107 (2005)
3. Bao-he, S.: Reliability and Optimal Inspection Policy of Inspected Systems. J. OR Transactions 11(1) (2007)
4. Bao-he, S.: Study on Optimal Inspection Polices Based on Reliability Indices. J. Engineering Mathematics 25(6) (2008)
5. Zong-shu, W.: Probability and Mathematical Statistics. China Higher Education Press (2008)
6. Mankiw, N.G.: Principles of Economics, International, 6th edn (2011)

The Research of Private Network Secure Interconnection Scheme in Large-Scaled Enterprises

Haijun Guo, Tao Tang, and Di Wu

China General Nuclear Power Group, ShenZhen, China
guohaijun@cgnpc.com.cn

Abstract. In the process of rapid development of large enterprises, some independent and closed business requirements arise, such as the core design, production data, security monitoring and so on.These businesses are required for carrying on the private network seprarated from the enterprise office network, but it can not be completely isolated because it exists business integration needs and economic demands such as saving investment, simple operation, so it brings a contradiction between the private network construction and cost control. In this paper, combining with the firewall technology and the optimization network framework, we give the basic ideas and implementation solutions to this contradiction.

Keywords: Private Network, Integration requirements, Logic isolation, Firewall, ACL, Two layer design.

1 Introduction

In the process of rapid development of large enterprises, some independent and closed business requirements arise, such as the core design, production data, security monitoring and so on. These businesses are required for carrying on the private network separated from the enterprise office network, but it cannot be completely isolated because it exists business integration needs and economic demands such as saving investment, simple operation. So it brings a contradiction between the private network construction and cost control. In this paper, combining with the firewall technology and the optimization network framework, we will research a construction scheme of the private network which satisfies business requirements, security requirements, and cost saving. Refering to this thesis, network administrators can deploy safety and economy private network more easily.

2 Example for Private Network

Taking the typical cases of a network equipment manufacturersthe--core design private network of a large enterprise as example, We will analyze the secure interconnection scheme. The core design data belongs to the intellectual property rights of the enterprise, with high security and confidentiality, and needs to transfer in the private network. The core private network design is built like the figure 1.

H. Sun et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 329, DOI: 10.1007/978-3-319-12286-1_43

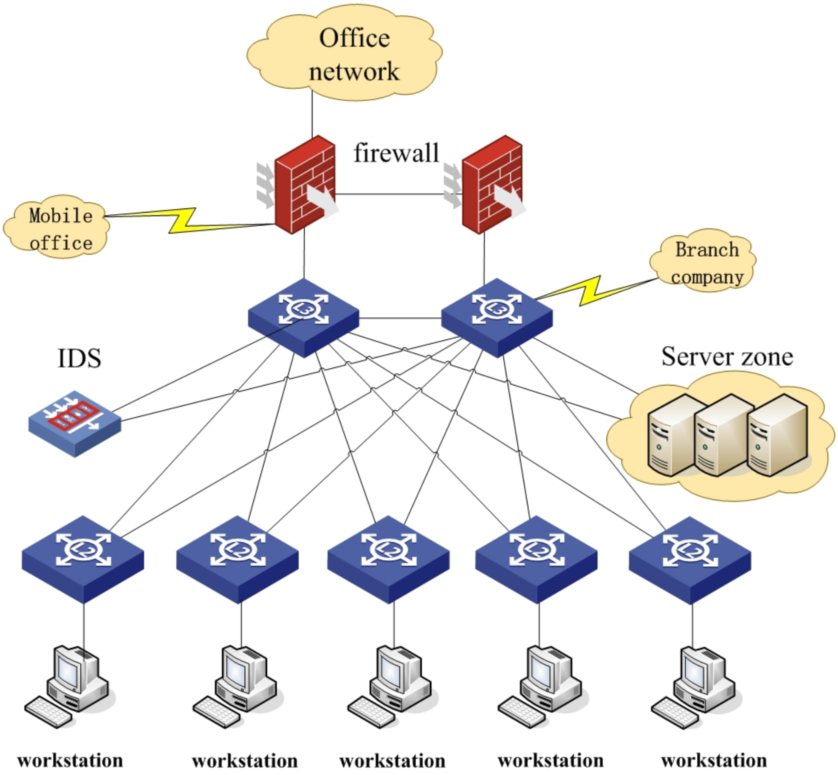


Fig. 1. Example for private network

In the aspect of network architecture, the aggregation switches are in hot standby situation, the access switch and aggregation switch are connected with redundant link, aggregation switch and the management network are isolated logically through the firewall. Mobile office users use VPN to access to design network through the Internet export connected with the firewall Branch company users access to design network through the digital line connected with the aggregation switch.

In the aspect of server deployment, several core design service servers such as design server, image server, test server, database server, the storage server as well as some security servers for access authentication, anti-virus, system upgrades and VPN access are deployed in the internal network.

In the aspect of safety control, the data between private network and office network are isolated through the firewall policy, at the same time, firewall opens design network equipment monitoring service port. Network workstations install antivirus software and copy protection software, also enable access authentication to prevent illegal access from unauthorized computer. Finally, IDS is deployed to monitor internal data flow.

3 General Scheme Design

According to the core design private network of a large enterprise, we can refine the general design method of Private Network secure interconnection scheme in large-scaled enterprises, including network architecture, system deployment, security control etc...

3.1 Design Points

First, recommend the network architecture apply two layers with access and core because of its small-scale, dual link redundancy design of the key equipment, a new and independent network segment to facilitate the implementation of special network security strategy, and the spanning tree protocol to prevent the two layer network loop enabled in the private Network.

Second, following the international general standards, the brand and model of private network equipment uses advanced and mature technology, to ensure the network with compatibility, high stability, easy scalability, and manageability.

Third, according to the information security requirements, security system of private network includes: network boundary firewall to separate office information, network intrusion detection system to monitor the internal network attack behavior, terminal security access, antivirus and anti-data copying system; remote computer through Internet with VPN or Citrix system to access the private network, remote branches access with digital line.

Fourth, internal servers of private network in the intranet server zone, the internal servers which interact information with external deployed in DMZ District, internal computers access to office network servers through the firewall security policy.

Fifth, large enterprise generally has a centralized network monitoring platform in office network; private network equipment should be included in the same platform for monitoring, to avoid repeated construction of network monitoring platform.

3.2 Detailed Solutions

Network architecture---Private network is an extension of the enterprise office network, as the bearer of a special platform for enterprise business; it has high independence and security. The private network is logically isolated from external network through firewall, and makes up internal construction according to the actual business needs as the figure 2.The architecture description of figure 2:

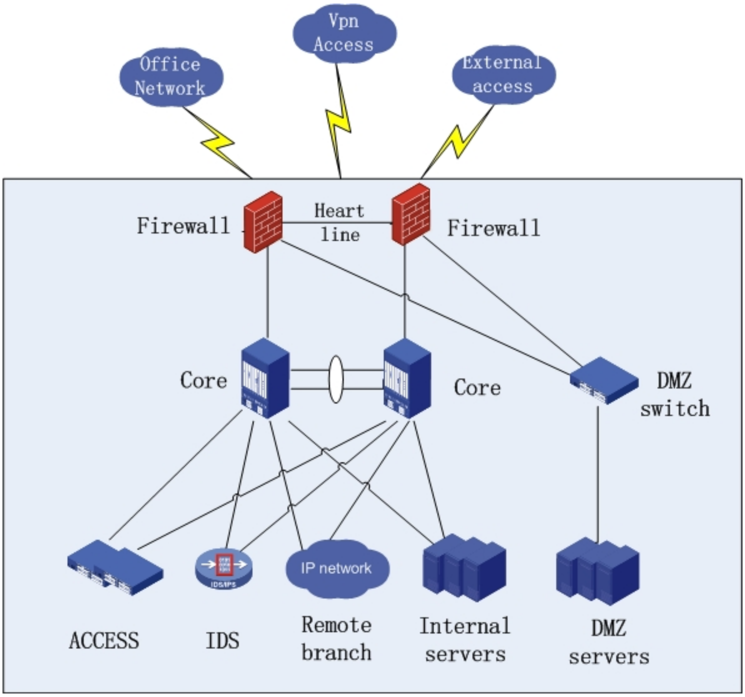


Fig. 2. General scheme for private network

(1)The key equipment such as perimeter firewalls and core switches have hot standby deployment, with dual-link interconnection.

(2)Access switches and internal business servers connect directly with the core switches.

(3)Remote branches connect directly with the core switches if they are needed.

(4)Personal office PC with access need to private networks via the Internet with VPN or Citrix system.

(5)If it exsits external unit access requirements, visited business servers are deployed in the DMZ area of the private network, and accessed through the firewall.

(6)The private network with a high security level should be deployed IDS devices in the core switches.

(7)All network equipment and terminals use static IP addresses, and the IP address must be independent of office network.

(8)Security Policy---By default, Firewalls discard the all packet from source IP addresses to destination IP address, you need to set security policies for individual permitted office network servers through the firewall , including domain controller servers, network management servers, patch upgrade servers, viruses upgrade servers, other servers for business demand and so on.

3.3 Simplified Scheme

If a private network has no external or remote access needs, we can use the datacenter firewall and Access Control List to achieve isolation with office network and special network, the private network shares local area network core switch, server firewall, server core switches and IDS equipment of the office network, such as the specific architecture of figure 3.

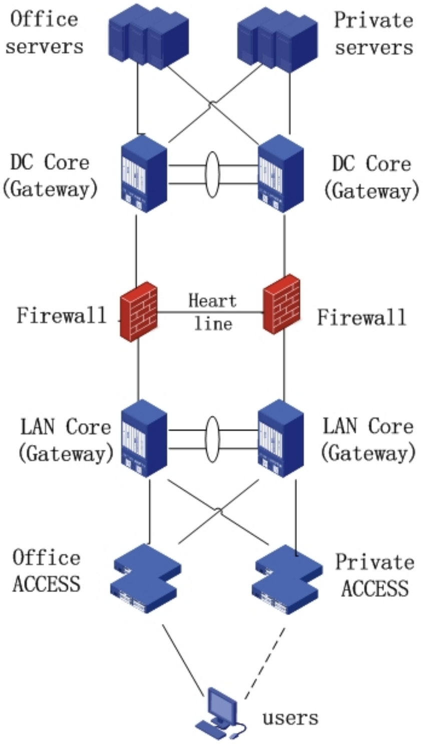


Fig. 3. Simplified scheme for private network

The security policy is set as follows:

(1) Conduct ACL strategy in the LAN access gateway to isolate office network user segment（X.0.0.0/8）from private network user segment (X.X.0.0/16)，like this:

```
acl number 3000
rule 0 deny tcp source X.0.0.0 0.255.255.255 destination X.X.0.0 0.0.255.255
rule 5 permit tcp source X.X.0.0 0.0.255.255 destination X.X.0.0 0.0.255.255
```

(2) Conduct ACL strategy on the firewall to isolate office network server from private network user, to interconnect the office network user and office network server, to interconnect the private network user and private network server;

(3) Conduct policies on the server gateway to isolate the office network server and private network server;

(4) In order to facilitate the implement strategy plan, we need private network segment, general office network segment is set to X.0.0.0/8, private network segment can be set to X.X.0.0/16, after the security policy is set, the data flow will be as shown the figure 4: office network users cannot access private network business system, private network users cannot access the office network business system, office network business system and private business system cannot mutually visit, office network user and private network users cannot mutually visit. (green arrows indicate the data flow, the red arrows represent data block)

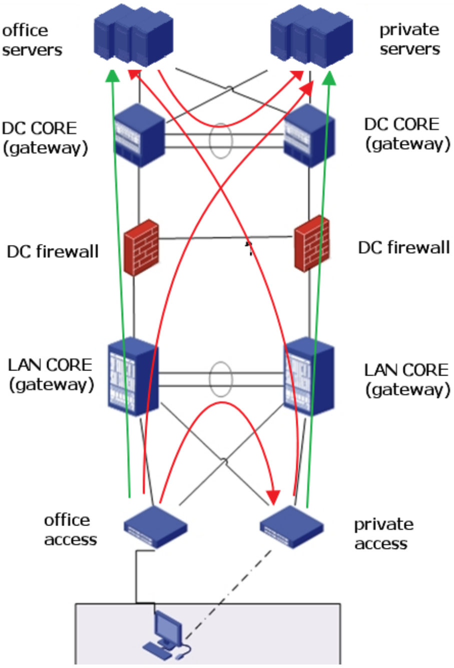


Fig. 4. The data flow of Simplified scheme

(5) Through the safety authentication and data encryption for network service system, the plaintext information of private network cannot flow into the office network, which solves the private business data security issues in common terminals.

Through the above strategy to achieve the control effect of office network and private network segment data isolation,at the same time, simplified private network saves the purchase of core switches, firewall, IDS, computer terminal investment, and only needs to purchase the access switch and the special business server special network, which greatly reduces the investment in infrastructure, and avoids the waste of resources.

3.4 Related Technology

(1) Firewall technology--Firewall is a kind of isolation control technique, through scanning all internal network and external network flow via the firewall, to carry out

data stream transmission or discard in accordance with security strategy matching rules. After the Firewall is deployed, external network will not be able to access the internal network server, but some servers has the access needs for external, in order to solve this problem, the firewall has been set up a buffer zone between non safety system and safety system, in this area, some servers open to the external network can be placed, this buffer zone is called DMZ. Firewall forbids the external network area access the inner area directly, but DMZ can communicate with the external network area, also can communicate with the internal network area with the limit of the safety rules.

(2) Intrusion detection system--Intrusion detection system (IDS) is a kind of network security devices which can monitor network transmission in real time, sound an alarm once detect the suspicious transmission or take active response measures. According to a certain security policy, IDS monitors the network and system operation conditions, tries to find the attack attempts, aggression or attack results, to ensure that the confidentiality, integrity and availability of the network system resources. Different from firewall, IDS is a listening device, not across any link, and works without network traffic flows through.

(3) Virtual private network--Virtual private network (VPN) is a kind of remote access technology to solve the problems that how the field trip user safely access intranet server. Field staff in local connects to the VPN server via the Internet, and then through the VPN server into the enterprise intranet. In order to ensure data security, data communication between VPN server and client are encrypted. With data encryption, you may consider data transmits safely in a private data link, which are like a private network, but VPN actually uses the internet common link.

4 Conclusion

In this paper, using the firewall technology and the optimization network framework, we design the private network secure interconnection scheme in large-scaled enterprises. The benefits and advantages of the private network model are both meeting the special businesses, and avoiding the waste of investment.It provides a reference scheme for the network construction of large enterprises.

References

[1] Chun, Z.: Construction of enterprise information security system based on ISO27001 vision. Journal of Information (May 2009)

[2] Zhang, J.: Network security information management system based on ISO27000. Nanjing Agricultural University (2009)

[3] Zhou, G.: The development research of intrusion detection system evaluation and technology. Modern Electronic Technology (December 2004)

[4] Zhan, G.: The method research of special network security in the construction of. Sichuan University (2002)

[5] Niu, W., Xie, H.: Application of private network technology in the enterprise MES network design. In: The 2009 Annual Meeting Proceedings of China Measurement Association Branch of Metallurgy (2009)

[6] Zhao, L.: The application of virtual private network VPN in the resource sharing in library. Modern Library and Information Technology (February 2005)

[7] Yi, Z.: Special network model design and implementation with IP centralized management function. The Journal of Guangxi University For Nationalities (January 2014)

[8] Li, H.: A solution of the enterprise special network security. Journal of Zhanjiang Normal University (March 2002)

[9] Yu, L.: Firewall in network security. hope monthly (first half) (December 2007)

Author Index